Master's thesis

Master's Programme in Data Science

Automated Grading of Newborn EEG Background Activity

Ilse Tse

May 15, 2019

Supervisor(s): Dr. Jukka Kohonen
Dr. Teemu Roos
Dr. Sampsa Vanhatalo

Examiner(s): Dr. Jukka Kohonen
Dr. Teemu Roos

University of Helsinki
Faculty of Science
P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Degree programme
Faculty of Science	Master's Programme in Data Science

Tekijä — Författare — Author

Ilse Tse

Työn nimi — Arbetets titel — Title

Automated Grading of Newborn EEG Background Activity

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidantal — Number of pages
Master's thesis	May 15, 2019	59

Tiivistelmä — Referat — Abstract

Background Electroencephalography (EEG) depicts electrical activity in the brain, and can be used in clinical practice to monitor brain function. In neonatal care, physicians can use continuous bedside EEG monitoring to determine the cerebral recovery of newborns who have suffered birth asphyxia, which creates a need for frequent, accurate interpretation of the signals over a period of monitoring. An automated grading system can aid physicians in the Neonatal Intensive Care Unit by automatically distinguishing between different grades of abnormality in the neonatal EEG background activity patterns.

Methods This thesis describes using support vector machine as a base classifier to classify seven grades of EEG background pattern abnormality in data provided by the BAby Brain Activity (BABA) Center in Helsinki. We are particularly interested in reconciling the manual grading of EEG signals by independent graders, and we analyze the inter-rater variability of EEG graders by building the classifier using selected epochs graded in consensus compared to a classifier using full-duration recordings.

Results The inter-rater agreement score between the two graders was κ=0.45, which indicated moderate agreement between the EEG grades. The most common grade of EEG abnormality was grade 0 (continuous), which made up 63% of the epochs graded in consensus. We first trained two baseline reference models using the full-duration recording and labels of the two graders, which achieved 71% and 57% accuracy. We achieved 82% overall accuracy in classifying selected patterns graded in consensus into seven grades using a multi-class classifier, though this model did not outperform the two baseline models when evaluated with the respective graders' labels. In addition, we achieved 67% accuracy in classifying all patterns from the full-duration recording using a multi-label classifier.

ACM Computing Classification System (CCS):
Applied computing → Life and medical sciences → Health Informatics

Avainsanat — Nyckelord — Keywords

EEG, AGS, brain monitoring, neonate, multi-class classification, multi-label classification, SVM

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

Medical Terminology

Anoxia	hypoxia to the severity resulting in permanent damage
Apgar score	an index used to evaluate the condition of a newborn on a scale from 0 – 2 for five categories (respiration, heart rate, muscle tone, reflexes, skin color), summed so a 10 is the perfect score
Artifact (in EEG)	extraneous noise with sources other than the brain
Asphyxia	lack of oxygen in the body, leading to unconsciousness or death.
Cerebral palsy	a neurologic disorder characterized by muscle incoordination and speech disturbances
Electroencephalography (EEG)	a brain imaging method that captures electrical activity in the brain and amplifies them at the scalp with conductive media and electrodes
Encephalopathy	a brain injury
Hypothermia	lowering of body temperature to subnormal degree
Hypoxic	when there is inadequate oxygenated-blood supply
Intrapartum	during the act of birth
Ischemic	when there is inadequate blood flow to the brain
Neonate	newborn less than 4 weeks of age
Perinatal	the period around birth
Seizure	convulsions resulting from abnormal discharges of electrical activity in the brain

List of Abbreviations

ACNS	American Clinical Neurophysiology Society
aEEG	amplitude-integrated EEG
AGS	automated grading system
BSP	burst-suppression pattern
cEEG	continuous EEG
CV	cross validation
ECOC	error-correcting output code
EEG	electroencephalography
HFF	high-frequency filter
HIE	hypoxic-ischemic encephalopathy
IBI	inter-burst interval
ICU	intensive care unit
LFF	low-frequency filter
LOO-CV	leave-one-out cross validation
NICU	neonatal intensive care unit
qEEG	quantitative EEG
RBF	radial basis function
SWC	sleep-wake cycle
SVM	support vector machine

1. Introduction

In neonatal care, the non-invasive electroencephalography (EEG) provides a method to monitor the cerebral recovery of newborns who have suffered a brain injury. EEG is an electrobiological imaging tool that records electrical activity in the brain. The electrical activity is often categorized into background activity patterns that are associated with the health of the brain, where it may be a sign of cerebral dysfunction if the EEG background reveals abnormal patterns of discontinuity. Such analysis is often based on the visual interpretation of changes in the rhythms and patterns of background activity [16] [31, p. 3].

Unfortunately, the expertise required to interpret the signals reliably is often unavailable or limited at many Neonatal Intensive Care Units (NICUs), which motivates the need for an automated grading system that can accurately and reliably detect patterns from the EEG features. This can ease the burden of a busy NICU by aiding any attending physician that may not have the EEG background to interpret the activity.

Methods for an automated grading system to grade EEG background abnormality have been developed by various groups [17] [21] [32] [1][22][20]. The algorithmic pipeline of such a system typically begins with filtering and generating features from raw EEG signals, and then training the features in a classification model to discriminate between different patterns or states.

This thesis aims to address a part of this algorithmic pipeline, which entails classifying EEG features into grades of abnormality. We explore the relationship between inter-rater variability in the EEG grades and the classifier predictions, and we are particularly interested in reconciling the manual grading of EEG signals by independent graders. We do so by comparing separate classifiers using two graders' EEG grades as class labels to ascertain how reliable a classifier can be in predicting EEG grades when there is ambiguity in the truth of the class labels.

The rest of the thesis is organized as follows: Chapter 2 begins with a brief overview of the human brain, and describes the fundamentals of EEG. Chapter 3 details the use of EEG monitoring in newborns with hypoxic-ischemic encephalopathy and the associated background abnormality. Chapter 4 discusses related work in automated grading systems of EEG data, and Chapter 5 describes multi-class and multi-label classification theories.

The methodologies used for the experiments are discussed in Chapter 6, and Chapters 7 and 8 explain the results and analysis of the experiments. We also briefly discuss the implications of using different scoring systems and having inter-rater disagreements. Finally, Chapter 9 concludes this thesis.

2. The brain and electroencephalography

2.1 The brain

The brain can be divided into three primary regions: the cerebrum, cerebellum, and the brain stem. The cerebrum regulates movement, sensory awareness, and emotional and behavioral expression. The cerebellum regulates voluntary motor movements and maintains balance, and the brain stem regulates involuntary functions, including heart regulation and respiration [3, p. 183].

The cerebrum consists of the left and right hemisphere structures that are divided by a fissure. The cerebral cortex makes up the outer layer of gray matter, or large bodies of neuronal cells, in the cerebrum. It comprises specialized lobes that are responsible for sensory and motor functions (Figure 2.1). The frontal lobe is responsible for executive functions and primary motor control, and the parietal lobe is responsible for integrating sensory details. The temporal lobe is responsible for sensory details as well, particularly auditory and visual information, and memory. The occipital lobe is the dominant visual processing area [3, p. 205–211].

In the human brain, there are approximately 85 billion neurons of different types [3, p. 24]. Neurons are cells that exchange information over large distances in the human body by firing electrical signals. A neuron's structure can be described by its axon, soma, and dendrites [3, p. 49]. A class of neurons in the cerebral cortex is the pyramidal neuron. It is distinguished by its pyramidal shape formed by short clusters of basal dendrites at the bottom and longer apical dendrites at the top end of the soma (Figure 2.2). Pyramidal neurons make up $\frac{2}{3}$ of all neurons in the cerebral cortex, contributing to a considerable amount of electrical activity.

Electrical activity is produced when a neuron is activated and the synaptic excitation at the dendrites of the neuron produces a flood of local current flows, causing a difference of electrical potential in the interior of the neuron and in the extracellular space. Electrical activity can travel through many layers of non-neural tissue, skull, and skin to reach

electrodes placed on the scalp to capture the signal. Pyramidal neurons are of interest since they are near the cortical surface in large numbers and their electrical activity can be recorded at the scalp [3, p. 82–107, 647].

2.2 Electroencephalography

Electroencephalography (EEG) is a brain imaging method that captures electrical activity in the brain and amplifies them at the scalp with conductive media and electrodes. EEG gives a limited view of the cerebral cortex since the electrical activity needs to travel some distance and penetrate many layers in order to reach the electrodes. EEG can only detect the neuron potentials if the neurons synchronously fire along with neighboring neurons in a minimum area of 6–10 cm^2, limiting the spatial resolution to this scale. Moreover, EEG recordings that use a fewer number of electrodes describes a wider subcortical region than if more electrodes were used, which also limits the spatial resolution. However, EEG have an excellent temporal resolution at milliseconds [31, p. 3–5].

For adults and newborns, the standard placement of electrodes is determined according to the international 10-20 system, though infant electrode placement requires fewer electrodes to account for their smaller scalp (Figure 2.3 – 2.4). The 10-20 system ensures that the signals can be referenced to a source location and are standardized across different EEG systems by directing the placements on anatomical landmarks on the head. This also ensures the reliability of the signals despite differing brain structures due to individual variation [31, p. 4].

The anatomical landmarks for electrode placement are markers on the sagittal midline and coronal midline. The sagittal midline is defined from nasion, at the bridge of the nose, to inion, at the back of the head. The preauricular points by the ears define the coronal midline. The electrodes are then placed at 10% and 20% intervals along these markers.

The nomenclature for electrode placement is indicated by capital letters – F (Frontal), Fp (Frontal polar), C (Central), T (Temporal), O (Occipital), P (Parietal), A (Preauricular) – followed by either a numerical, **z**, or **p** suffix. The odd and even numbers indicate placement at the left and right hemisphere respectively, and larger numbers indicate a further placement distance from the sagittal midline. **z** indicates a location on the midline, and **p** indicates the frontal pole [31, p. 4].

Each electrode pair makes up a channel of an EEG trace and is represented by the difference in potentials between two electrodes. Two methods of combining the channels are the bipolar montage and the referential montage.

In the bipolar montage, the combination of channels can be arranged as a chain in either an anterior-to-posterior or transverse direction. When the first electrode in the

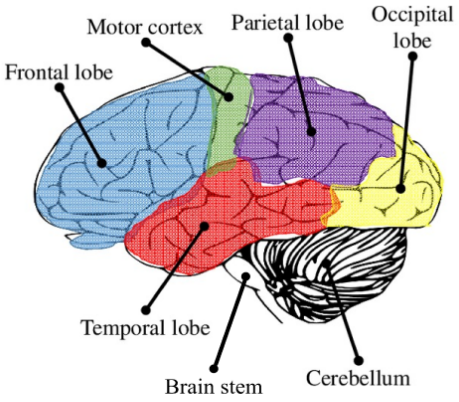

Figure 2.1

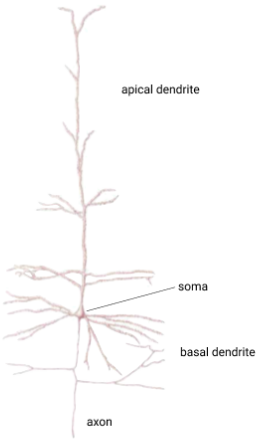

Figure 2.2

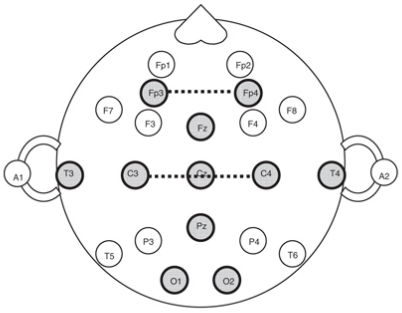

Figure 2.3

Figure 2.4

Figure 2.1: The lobes of the cerebral cortex (Modified from Jebelli, Houtan, Sungjoo Hwang, and SangHyun Lee, 2017). **Figure 2.2:** A cortical pyramidal cell, found in the cerebral cortex (Modified from Bear, 2007). **Figure 2.3:** EEG electrode placement for adults and infants (shaded regions) (Wusthoff, 2009). **Figure 2.4:** An infant wearing a **wave**guard^{TM} neonatal EEG cap that can record with 24 – 43 electrodes (Photo courtesy of Sampsa Vanhatalo).

chain captures a more positive amplitude than the second, the pair subtracts from each other, and a positive potential is displayed. Similarly, if the first electrode in the chain pair is more negative than the second, a negative potential is displayed.

This is distinguished from the referential montage, in which the combination of channels is the electrical potential difference between any given electrode and a pre-selected electrode. Whereas the bipolar montage can amplify local potentials, the referential montage can more closely represent the absolute potential of an electrode. When interpreting EEG patterns, one thing to keep in mind is that the activity captured do not reflect a single cerebral region, but rather of multiple [36].

2.2.1 Standard EEG review settings

The American Clinical Neurophysiology Society recommends to record with least 21 electrodes and displaying 16 channels for adults, and modified to 9 electrodes for neonates [7]. During an EEG recording, the channels are displayed as horizontal polygraphic outputs on a display monitor for visual interpretation by a neurophysiologist. A positive potential is indicated by a downward deflection, while a negative potential is indicated by an upward deflection. Visually, the EEG signals are in the form of waves, typically of a sinusoidal shape, and they reflect the spatial and temporal properties of the signal, such as amplitude and frequency [31, p. 19].

Standard EEG review settings determine the activity output's amplification, frequency filtering, and time scale that aid in the readability of the EEG record. Cerebral potentials are often low-amplitude, so differential amplification is used to filter out electrical noise by subtracting it by the uniform noise that appears identical at multiple electrodes.

Common settings for amplitudes are in the range of 5 to 10 μV/mm, though a wider range of settings are often used. Higher amplitude activity must be compressed to fit in the EEG display monitor and may hamper the visibility of lower amplitude activity, so a 2 μV/mm setting is often a practical amplification limit for cerebral activity. EEG devices often record frequencies in the range of 0.1 to 125 Hz, but standard review settings use a narrow bandpass filter of a low-frequency filter (LFF) and a high-frequency filter (HFF) to output activity in the range of 1 to 70 Hz. A 1-Hz LFF is equated to a 0.16 s time constant, and has a negative linear relationship; as LFF increases, the time constant decreases. [31, p. 4 – 7]. In neonatal EEG recordings, the LFF is often set to record slower frequency activity in the range of 0.005 to 0.01 Hz or 0.5 Hz [5, p. 21]. For infants, EEG activity over 32 Hz is negligible [32].

Since electrical activity occurs many layers below the scalp, the EEG signals that are captured often have noise, or artifacts, that must be filtered out. Noise in the EEG record may be from biological or non-biological sources. A common source of non-biological noise is in the electrical current in the environment, so the notch filter is set to the power supply's AC current to reduce this noise. Common sources of biological artifacts include muscle, eye, heart, or respiratory movements and may appear as high-frequency noise, which can be reduced by lowering the HFF setting or increasing the LFF setting.

The time scale of the activity determines the expansion or compression of the activity on the display; for instance, greater horizontal compression may aid in observing slow seizures evolution. The recommended time scale is 10 s per page, or 12 s per page for wider screens. [31, p. 7].

The conventional electroencephalogram (cEEG) uses a full array of electrodes to

capture cerebral electrical activity, while the amplitude-integrated electroencephalogram (aEEG) is a simpler method of monitoring the brain by displaying a bandpass-filtered, time-compressed EEG from one channel as it typically records from only 2 electrodes. While cEEG is a robust method of EEG monitoring, it is a resource-expensive procedure in the NICU as it takes a specialized neonatal neurophysiologist to perform a multi-channel EEG recording on newborns, filter out artifact sources, and to interpret the background activity. As a result, the use of aEEG to monitor neonatal EEG is rising in popularity in NICUs. Although there are limitations to aEEG due to the reduced number of electrodes used in recording, it displays EEG background activity trends that are useful determining the severity of EEG abnormality in neonates [28].

2.3 Identifying EEG background activity patterns by their features

The goal of analyzing EEG features is to associate them with EEG background activity patterns with clinical significance. EEG features are in the form of sinusoidal waves, often categorized by their rhythm, which describes activity with waves of relatively constant period. Background activity refers to any EEG activity outside of a distinguishable pattern that an interpreter is currently observing. The main terminology used in assessing EEG patterns are those indicative of pattern location and type, with modifiers that describe the persistence, duration, amplitude, frequency, and sharpness of the pattern [19].

To identify EEG patterns by their features, the current segment or epoch must be categorized as either a transient, attenuation, or repetition. A transient describes an isolated wave that is distinguished from background activity; attenuation refers to a reduced EEG amplitude; repetition refers to recurrence of the transient several times without interruption by background activity. Then, the feature can be categorized further by the distribution of the wave's electrical field at the scalp – focal, distribution to one electrode and its immediate neighbors; hemispheric, distribution to electrodes on a unilateral, anterior and posterior to the coronal midline; bilateral, distribution to electrodes on both sides of sagittal midline, limited on either anterior or posterior of coronal midline; generalized, distribution to electrodes on sagittal and coronal midline [31, p. 20–22].

Apart from using the morphological features to describe the EEG activity, neurophysiologists can also perform quantitative EEG (qEEG) analysis to extract additional features that complement the features used in visual interpretation [14]. A toolbox for neonatal qEEG analysis is available for open-source use [34].

2.3.1 Visual interpretation of neonatal EEG patterns and inter-rater reliability

Visually interpreting neonatal EEG background activity can be a challenging task. To monitor the brain recovery of a newborn, there is a need to frequently assess the continuous EEG signal recorded over many hours in order to evaluate how the EEG background patterns are evolving, making the EEG interpretation a time-consuming task. Despite the use of aEEG that simplifies the EEG display, interpreting the signals is still technically demanding due to the individual variations seen in patients. Moreover, there is the additional challenge of subjective interpretation of the EEG signals, especially if assessing the EEG involves multiple independent graders.

When EEG patterns are visually graded by more than one electroencephalographer, the inter-rater reliability between the graders indicates to what degree is the interpretation of the patterns reliable and reproducible. High inter-rater agreement scores may indicate the EEG features' good reliability and reproducibility due to agreement when describing the patterns. Standard metrics that can be used to evaluate the inter-rater agreement in subjective scoring tasks include Cohen's kappa and Fleiss' kappa.

Cohen's kappa statistic (κ) is used to evaluate the inter-rater agreement between two raters while taking into account of chance (dis)agreements for categorical variables. When there are more than two raters, Fleiss' kappa may be used. A κ of 1 means perfect agreement, and κ of 0 means the variations occurred by chance [38].

Given two observer's (dis)agreement information (Table 2.1), Cohen's kappa statistic can be calculated with equations (2.2), (2.1), and (2.3). If a particular class agreement is more important than another, the kappa can also be weighted to reflect this bias [38].

		Observer 1 results		
		Yes	**No**	**Total**
Observer 2 results	**Yes**	a	b	m_1
	No	c	d	m_0
	Total	n_1	n_0	n

Table 2.1: Results table from two observers for a two-category (Yes/No) dataset. a and d are observations in agreement, b and c are numbers of observations in disagreement (Modified from Viera, 2005).

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2.1}$$

$$p_o = \frac{a + d}{n} \tag{2.2}$$

$$p_e = \left[\frac{n_1}{n}\frac{m_1}{n}\right] + \left[\frac{n_0}{n}\frac{m_0}{n}\right] \tag{2.3}$$

p_o: observed agreement, p_e: expected agreement

A κ value of < 0 indicates less than chance agreement, 0 – 0.20 indicates poor inter-rater agreement, 0.21 – 0.40 indicates moderate agreement, 0.61 – 0.80 indicates good agreement, and 0.81 – 1.0 indicates excellent agreement [28].

The kappa may not be representative of a low overall agreement if the observations are rare since the statistic takes into consideration the prevalence of the observations. A solution to this misrepresentation is to observe the agreement on the individual class labels [38].

3. The analysis of EEG background pattern abnormality

Brain development from the neonatal period to the first year of life is marked by the rapid growth of cortical hemispheres; gray matter, or neuronal bodies, may increase by 149%. During this period of rapid growth, the brain may be more vulnerable to injuries that could disrupt its typical development of brain structure and function [13]. When a newborn suffers a brain insult, brain monitoring tools such as the electroencephalography (EEG) can give insights to the cerebral function of the developing brain. These infants display characteristic patterns of EEG abnormality that can help physicians monitor their cerebral recovery or deterioration, and estimate the severity of the injury.

3.1 The use of EEG monitoring to assess hypoxic-ischemic encephalopathy in neonates

Hypoxic-ischemic encephalopathy (HIE) is a brain injury that occurs when there is an interrupted flow of oxygenated blood to the brain. Perinatal HIE affects approximately 1.5/1000 full-term neonates worldwide, and has approximately 15–20% mortality rate in newborns. 25% of HIE survivors exhibit long-term neurological deficits, such as cerebral palsy, learning disability, and epilepsy [15].

HIE is an evolving injury in that hours after the initial trauma, secondary injuries including neuronal necrosis and apoptosis may occur. While newborns with mild HIE typically recover within 24 hours, those with moderate to severe encephalopathy have an increased chance of developing seizures [16].

A common therapy includes induced hypothermia by lowering the cerebral temperature to 33 – 34 °C for 72 hours, and provides an opportunity to rescue some cerebral tissue and prevent neuronal death. This intervention is most effective for newborns with severe HIE if administered within a short time window of 2–6 hours after birth. Newborns with moderate HIE also benefit from therapeutic hypothermia if administered 6–12 hours after birth [9]. Moreover, HIE may periodically lapse into the reperfusion phase at 6–24

hours after the insult, during which interventions may be more effective [39].

To monitor the effects of therapies and also the brain recovery of newborns post-insult, EEG is a well-suited neuroimaging method since it is non-invasive and allows for continuous monitoring of cerebral activity. It is an effective tool in correlating cerebral recovery; EEG voltage at 6 hours after birth and EEG patterns at 3–6 hours after birth both correlates well with neurodevelopmental outcome of preschool age children who had suffered HIE at birth [27] [10]. Since the clinical symptoms of HIE appears early after birth, it is useful to start the EEG record as soon as possible, or within 10 – 48 hours after birth, for at least the first three days [16].

When visually interpreting the neonatal EEGs, information such as the age of the newborn, the clinical condition of the neonate such as the Apgar score, frequency of seizures, and medication taken, all create variability in the interpretation. Particularly for neonates who have suffered a hypoxic insult, a common treatment following epileptic seizures is the prescription of anti-convulsants, such as phenobarbitals. This may affect the EEG activity, so it is recommended to start the EEG record 30 minutes – 2.5 hours after administering the medication [16].

3.2 EEG background activity patterns

A major goal from EEG monitoring is to associate the background patterns with the clinical state of the brain. In the healthy newborn, a typical sleep-wake cycle (SWC) is a continuous transition between stages of EEG frequency band rhythms – the delta band (1–4 Hz, high amplitude), theta band (4–8 Hz), alpha band (8–12 Hz), beta band (13–25 Hz), and gamma band (> 25 Hz) (Table 3.1). The EEG signals differ depending on the state of alertness of the subject being measured. Full-term neonates typically spend 60% of their time in REM sleep, during which the EEG pattern is similar to the theta and alpha bands [24, p. 228–229]. Their mean duration of a SWC is 50 – 60 minutes, and the cycle repeats until it is interrupted by wakefulness every four hours [16].

In neonates who have suffered a hypoxic-ischemic insult, their background EEG activity tends to be abnormal due to the slowing of activity, amplitude depression, and increasing discontinuity with bursts, and the SWC will appear disrupted by discontinuity and bursts [16]. Discontinuity typically refers to low-amplitude activities with a duration of a few seconds, with interruptions by high-amplitude activities [24, p. 220]. For specifically categorizing the continuity of the background EEG, the American Clinical Neurophysiology Society (ACNS) suggests identifying the patterns as continuous (including flat or continuously suppressed patterns), brief periods of attenuation, discontinuous, and burst-suppression [19]. The specific definitions of EEG abnormal background activity patterns are diverse and are detailed further in section 3.3.

Sleep stage	Pattern	Description
Awake	Pattern 3: active moyenne (continuous) Diffuse low-voltage, irregular theta and delta	-
Non-REM Stage I - IV	Pattern 1: trace alternant a. Bursts of 4 – 5 s of 1 – 3 Hz b. Low-voltage activity, 4 – 5 s Pattern 4: continuous high-amplitude slow waves	1. Regular breathing 2. No eye and limb movements 3. Tonic neck EMG
REM	Pattern 2: Mixed low-voltage irregular, 2 – 4 Hz slow waves	1. Irregular breathing 2. Frequent eye and limb movements 3. No tonic neck EMG

Table 3.1: Sleep-wake cycling EEG pattern in full-term neonates (Modified from Nunez, 1995).

The degrees of slowing and discontinuity correspond to the level of severity of HIE, which can be described with three grades – mild, moderate, and severe. The clinical grade of HIE is associated with the EEG state of a newborn over a period of monitoring [39] [16]. Mild cases of HIE have subtle slowing and discontinuity of the EEG, and neonates with moderate to severe cases of HIE often display more abnormal activity with transient EEG depression. In severe cases of HIE, the activity may present themselves as burst-suppression, low-voltage, or even isoelectric (flat with zero difference of electric potential).

A distinctive feature of EEG discontinuity is the burst-suppression pattern (BSP), which is marked by the contrasting amplitude of the burst and suppression (Figure 3.1). Burst activity amplitudes can range from low to high, but suppression activity amplitudes range from inactivity to medium amplitudes. The difference between burst and suppression must be clearly visible at $\geq 50\%$ decrease in amplitude, though a definitive threshold is difficult to define due to the wide range of amplitude of the bursts [31, p. 151 - 152]. A common way to describe BSP is with the inter-burst interval (IBI), which describes the ratio of burst and suppression. IBIs tend to be shorter (<10 s) with higher amplitude [21].

The evolving nature of HIE means the corresponding EEG states are constantly evolving and are non-stationary, and may even be affected by external factors. The EEG background patterns may appear highly depressed if the newborn had been administered high dosages of anti-convulsant medication used to treat epileptic seizures following the insult. As the dosage increases, the ensuing EEG frequency slows into the theta and delta bands until it appears isoelectric [24, p. 232] [16]. In addition, the treatment of hypothermia may increase the EEG discontinuity of the neonate aged 3 – 6 hours, though this effect recedes with time [39].

The evolution from a discontinuous state to a more continuous, coherent form is a

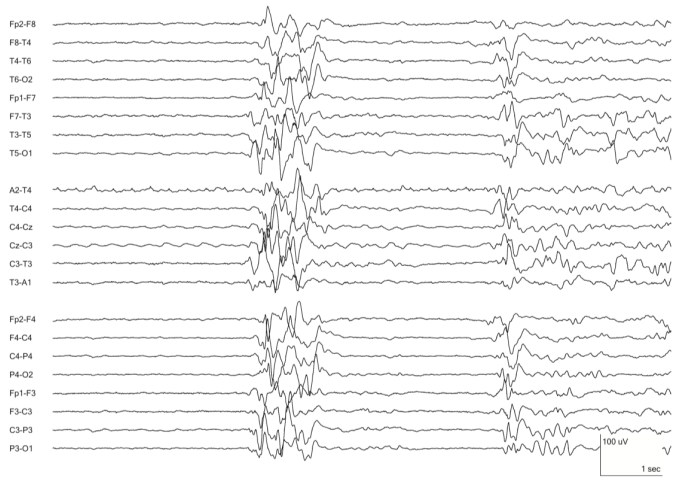

FIGURE 8-1 **Burst-suppression Pattern** • High amplitude sharp waves with intermixed spikes arise from relatively low amplitude background. The low amplitude suppressions comprise about half of the segment and the bursts vary in waveform components and duration. The EEG was recorded from a 2-year-old child with hypoxic-ischemic coma and multi-organ failure related to severe cardiomyopathy. (LFF 1 Hz, HFF 70 Hz)

Figure 3.1: Burst-suppression patterns (Stern, 2005).

developmental milestone towards a normal EEG state. EEG coherence is associated with a high chance of normal cerebral development, and the presence of normal sleep-wake cycling is also indicative of a favorable outcome. Ideally, the brain should recover from inactive/poor activity to continuous activity at the first 12 – 24 hours of life for a better chance of a normal outcome [39].

3.3 EEG classification schemes

There are currently no standardized values of amplitude and duration for assessing the background patterns, but multiple nomenclature have been proposed, and different studies have used their own definitions of EEG abnormality. Watanabe et al. (1980), Pressler et al. (2001), and Murray et al. (2009) all conducted different studies to evaluate the prognostic value of grading EEG abnormality in newborns with HIE [40] [26] [23].

Watanabe et al. took a 13 – 17 channel EEG recording of 132 full-term newborns with perinatal hypoxia. A recording was captured weekly for three weeks, and each record had a duration of 2 – 3 hours. The EEG were graded according to the classification scheme in Table 3.2. The classification showed high correlation between the EEG in the first week with the outcome at later years (the children in the follow-up had a mean age of 4.3 years). All infants who had normal EEG in the first week had normal outcome, and of the infants who had normal EEG in the second and third week, 72% in the second week and 43% in the third week had a favorable outcome. The authors also found high correlation between increasing levels of EEG abnormality and sleep-stage disturbances [40].

Pressler et al. (2001) assessed nine full-term infants with HIE with a 16-channel

video-EEG recording within eight hours of age (mean 5.6 hours). Each record had a duration of at least 60 minutes, and hour-long epochs were graded according to Table 3.2. Surviving infants were assessed for their neurodevelopmental outcome at 1 year. Two infants with moderate abnormality had normal outcome, and three who had inactive EEG grades but recovered within 12 – 24 hours had favorable outcome [26].

Murray et al. examined the predictive value of abnormal EEG results by taking the EEG recordings of 44 infants who had been diagnosed with HIE. The EEG recordings were taken within 6 – 72 hours from birth for a duration of 24 – 72 hours. The definitions in Table 3.2 were used to assess the EEG by hour-long epochs. The evaluation of the EEG grades using the classification scheme highly correlated with the outcome of the infants upon a follow-up at 24 months. The timing of the EEG recording affected the predictive value of the EEG grade; for EEG recorded at 6, 12, or 24 hours, normal/mildly abnormal EEG grades correlated with normal outcome (100% positive predictive value (PPV), 67 – 76% negative predictive value (NPV)). However, for EEG recorded at 48 hours, some of the moderate/severe EEG grades improved in EEG state as well as outcome (71% PPV, 93% NPV). For all cases, at 6 hours of age, normal EEG grades were associated with normal outcome, and very low-amplitude EEG grades were associated with abnormal outcome. For infants who had a maximal IBI of > 30 s or isoelectric trace and whose EEG did not recover by 12 hours, all had either severe neurological deficits or died [23].

From the three studies, it can be concluded that EEG grades of abnormality have high prognostic value if the EEG recording were done soon after birth at 6 – 24 hours of age. Normal grades usually correlated with good outcome, and cerebral recovery within 12 hours is also highly associated with favorable outcome.

Other studies have used different classification schemes to review the prognostic value of EEG grades of abnormality. A review by Walsh et al. (2011) compared 16 studies to consolidate general findings and interpret the common cEEG features (Table 3.2) used in the classification schemes. All studies used amplitude voltage and continuity patterns to describe the EEG abnormality, and most included descriptions of frequency, symmetry, SWC, transient, seizure, and maturity. Due to the wide range of the ages of the newborns, the follow-up duration, and differing outcome measures, it was difficult to make a fair comparison between the studies reviewed. However, there was general agreement in that the correlation between the clinical grade of HIE severity and the levels of EEG background abnormality was highest at extremes, and similarly so for outcome. Normal/mild EEG grades had high correlation with a mild HIE grade and good outcome, and severely abnormal EEG grades had high correlation with a severe HIE grade and poor outcome. Moderate abnormalities in EEG typically have low prognostic value in predicting the outcome, possibly due to the more diverse categorization of moderately abnormal injuries. Infants in this group are shown to benefit most from therapeutic hypothermia

[39].

The use of aEEG in NICUs has increased in popularity due to its compressed output for display, which reduces the number of channels to one or two for interpretation. However, this also reduces the amount of EEG information available to analyze, prompting studies to explore its limitations and to determine how to interpret the compressed activity for evaluating EEG background abnormality. Two widely accepted aEEG background activity scoring systems were proposed by Al Naqueeb et al. (1999), and by Hellström-Westas et al. (2006) [28].

Al Naqueeb et al's scoring system showed a close relationship between aEEG and the outcome of neonatal encephalopathy. Fifty-six infants at risk of encephalopathy were studied, and 40 were suspected of having suffered birth asphyxia. EEG recordings were taken with aEEG monitors for a median of 15 hours, and only the most abnormal trace sections were analyzed. For each subject, at least 30 minutes of recording were analyzed. Any administered medication was noted, and subsequent EEG recordings up to 30 minutes after administered anti-convulsants were excluded.

Two pediatric residents without extensive training in EEG were briefly trained in aEEG interpretation, and 50 EEG records were independently scored by an EEG expert and the two residents. The records were chosen to represent a broad spectrum of traces and contain records from normal infants from the control group as well. They also only contain traces that were recorded within 12 hours of birth to determine if EEGs can be used soon after birth. The categories of EEG patterns that were used are shown in Table 3.2.

The inter-rater variability showed good agreement of amplitudes (unweighted $\kappa = 0.85$). The overall agreement among the three graders was 0.85, and the two residents' agreement with the expert was 0.75 and 0.87 respectively. Subsequent follow up assessed the infant's outcome at 18 – 24 months of age. 19 of 21 infants who had been categorized as normal were normal at follow up, 13 of 15 infants whose EEG had been categorized as moderately abnormal/suppressed developed neurologic deficits or died, and no infants whose EEG had been categorized as suppressed had a normal neurodevelopmental outcome. This scoring system demonstrated high inter-rater agreement, as well as a high predictive value on prognosis, even with EEG taken close after birth at 5 hours [2].

A criticism of Al Naqueeb et al's simple scoring system is that the background baseline may vary and appear at higher margins in the presence of artifacts. This is especially true in abnormal tracings, and using this scoring system may risk underestimating the degree of abnormality in the trace.

Hellström-Westas et al. developed a more involved scoring system for categorizing abnormal aEEG signals and aims to be generalizable for use to categorize EEG of newborns of all ages, from very pre-term to full-term neonates (Table 3.2). This system lowers

the minimum amplitude margin below 5 μV and raises the maximum to above 25 μV.

Shellhaas et al's assessment of the two scoring systems, by Al Naqueeb et al. and by Hellström-Westas et al., aims to determine their inter-rater agreement in categorizing EEG background activity as well as the aEEG correlation with cEEG. 144 cEEGs were reviewed, and grades in consensus were chosen for conversion into aEEGs for experienced electroencephalographers to review using the two aEEG scoring systems. To convert cEEG into aEEGs, a single channel of the EEG was put through filters that attenuates to the amplitude levels of aEEG and also time compressed as such. To directly compare two scoring systems, Hellström-Westas et al's burst-suppression, low-continuous, and flat-tracing patterns were combined as to represent the suppressed (markedly abnormal) amplitudes [28].

The multi-rater κ score of 0.66 was good for Al Naqueeb et al's simple system, while Hellström-Westas et al's advanced system only had a moderate agreement score of $\kappa = 0.44$. Both systems had a similar fair agreement ($\kappa = 0.4$) with cEEG, which is considered the gold standard for categorizing EEG patterns. One reason for the low to moderate agreement scores may be due to the neonatologist graders' training since they worked at five separate institutions and represented three countries, and thus had different grade interpretation styles.

First Author	Grade / Description					
Watanabe (1980)	**Normal** Low voltage irregular (20 – 50 μV), mixed/medium voltage slow (30 – 100 μV), high voltage slow (50 – 150 μV), alternating tracing	**Minimally depressed** Discontinuous tracing – abnormally attenuated low voltage phase of tracé alternant	**Mildly depressed** Low voltage irregular, mixed/medium voltage slow, discontinuous tracing without high voltage slow	**Moderately depressed** Low voltage continuous patterns of low voltage irregular, poor activity (5 – 20 μV), discontinuous tracing without medium and high voltage slow	**Markedly depressed** Only discontinuous gracing with long flat periods or burst-suppression pattern	**Maximally depressed** Flat tracing (0 – 5 μV) Other: Low voltages consisting of low voltage continuous poor activity and low voltage irregular
Pressler (2001)	**Normal / mild abnormalities** Normal pattern for GA, incl. slightly abnormal activity, e.g. mild asymmetries, mild voltage depression	-	-	**Moderate abnormalities** Discontinuous activity with IBI of ≤10 s, other types of continuous activity, clear asymmetry or asynchrony	**Major abnormalities** IBI of 10–60 s, severe depression, no wake-sleep cycles	**Inactive EEG** Background activity of <10 μV, IBI of >60 s
Murray (2009)	**0** **Normal EEG** Continuous background pattern, ≥ 50 μV	-	**1** **Normal / mild abnormalities** Continuous background pattern with slightly abnormal activity	**2** **Moderate abnormalities** Discontinuous activity with IBI of <10 s, no clear SWC, or clear asymmetry or asynchrony	**3** **Major abnormalities** Discontinuous activity with IBI of 10 – 60 s, severe attenuation of background patterns, or no SWC	**4** **Inactive EEG** Background activity of <10 μV or severe discontinuity with IBI of >60 s
Walsh (2011)	-	-	**Mildly depressed amplitude** **Other names: Mild attenuation, mild voltage suppression, slightly decreased voltage** >25 μV, other ranges include 20 – 50 μV, 30 – 50 μV	-	**Low voltage** **Other names: Severe low voltage, marked generalized low voltage, poor voltage** <5 μV, other ranges include 5 – 15/20/25 μV, 10 – 50 μV, <20 μV, <30 μV	**Isoelectric** **Other names: Maximally depressed, electro-cerebral silence, inactive, monotonous low voltage, or flat** <10 μV, other ranges include <2/3/5 μV
Al Naqueeb (1999)	**Normal amplitude** lower margin >5 μV, upper margin >10 μV	-	-	**Moderately abnormal amplitude** lower margin ≤ 5 μV, upper margin >10 μV	-	**Suppressed amplitude** lower margin <5 μV, upper margin <10 μV, usually accompanied by BSP
Hellström-Westas (2006)				**Discontinuous** Discontinuous background with minimum amplitude variable, but below 5 μV, and maximum amplitude above 10 μV. **Burst-suppression** Discontinuous background with minimum amplitude without variability at 0 – 2 μV and bursts with amplitude >25 μV.	**Low voltage** Continuous background pattern of very low voltage (around or below 5 μV).	**Inactive, flat** Primarily inactive (isoelectric tracing) background below 5 μV

Table 3.2: Classification systems of neonatal EEG with abnormal background patterns. IBI indicates inter-burst interval, SWC indicates sleep-wake cycle. Al Naqueeb and Hellström-Westas' classification systems are used with aEEG output. The classification schemes vary by number of grades of abnormality, amplitude threshold and range, duration, and whether or not SWC is considered. Some grading systems use more qualitative terms to describe the patterns, while others are more elaborate from specifying the levels of abnormality to defining the amplitude range and duration.

4. Automated grading systems for neonatal EEG background abnormality

Automated methods for detecting EEG background abnormality have been developed to improve brain monitoring for neonates.

Löfhede et al. (2008) used Fisher's linear discriminant (FLD), a feed-forward artificial neural network (ANN), and a SVM to compare the algorithms' ability to distinguish bursts from suppression in the burst-suppression pattern. From the EEG recordings of six full-term newborns, 6 – 40 minute recordings of each newborn were used to ensure a minimum of ten bursts were included. An electroencephalographer identified burst-suppression patterns and then further categorized each pattern as burst or suppression. While the burst patterns have a higher amplitude than suppression, the entire burst-suppression pattern cannot be identified using an amplitude threshold. Five features, spectral edge frequency, 3 Hz power, median, variance, Shannon entropy, were calculated after pre-processing the raw EEG, and then normalized to the interval $[0, 1]$. Since the sample size was small, leave-one-out (LOO) cross-validation was used and six ROC curves were created from the validation sets to evaluate the classifiers. SVM was implemented using Matlab with an RBF kernel, though the parameters and weight parameters were undisclosed. SVM consistently outperformed both FLD and ANN, and FLD generally performed the worst [17].

Matić et al. (2012) built an automated grading system using a temporal profile to detect inter-burst intervals (IBI) to grade EEG discontinuity in neonates with mild HIE. EEG data from eight newborns with mild to moderate HIE were used, and three segments of 5-minute intervals from each recording were selected at the recording's beginning, middle, and end, and segments with artifacts were not discarded. The segments were visually graded by a neurophysiologist into four grades – normal (IBI $<$ 5 s), mild (5 s $<$ IBI $<$ 10 s), moderate (10 s $<$ IBI $<$ 5 s), and severe (IBI $>$ 20 s). Another neurologist then visually marked the beginning and end points of each IBI.

The EEG segments were generated with an adaptive segmentation algorithm to

segment the EEG channels based on two moving average windows that detected the largest difference in amplitude and frequency values between the windows as segment points, creating quasi-stationary epochs. The segments were further classified into low, medium, or high amplitude distribution categories to mimic human visual interpretation. The detection of IBIs is done by creating a temporal profile membership function for every class of IBI. The temporal profile function outputs a value which indicates the number of channels where the epoch falls under IBI definition. If the temporal profile signal exceeds a threshold of over half the total number of channels, the time of that signal defines the duration of the IBI. The algorithm correctly detected most IBIs, and most errors were in identifying low amplitude bursts that occurred in the beginning of EEG recordings, a few hours after the insult. While it could detect suppressions, it had trouble detecting transitions between bursts and suppressions [21].

Stevenson et al. (2013) have built an automated grading system using a multi-class linear classifier for classifying EEG abnormality for the neonatal population. 54 full-term neonates who had suffered from HIE had their EEGs recorded within 12 hours of birth for 12 – 72 hours using an 8-channel bipolar montage. Approximately one hour of EEG data were selected from each subject for analysis. The segments were chosen to ensure a relatively constant grade, and epochs with major artifacts were excluded. Epochs with minor artifacts were excluded from training but were included in the final evaluation as the model was expected to reject the artifacts automatically [32].

The EEG data was then independently graded by two neonatologists, and epochs in disagreement were subsequently reviewed and discussed by the graders until a consensus was reached. The inter-rater agreement was high (Cohen's $\kappa = 0.868$). The scoring system they used corresponded to grade 1, normal/mild abnormalities; grade 2, moderate abnormalities; grade 3, major abnormalities; grade 4, inactive.

The automated grading system pipeline involved pre-processing the EEG by normalizing the features using a Box-Cox transformation to ensure similar statistical distribution across all features. Then, they performed feature extraction to generate sub-signals before classifying the data using a multi-class linear discriminant classifier, using Cohen's kappa as a loss function. During training, they first expanded the outputs into six grades by creating two additional transition states to account for the natural cycling between grades 1 and 2 during sleep. They then combined the expanded states into the original four during post-processing.

Originally a binary classifier, the linear discriminant classifier was modified for the multi-class problem by decomposing into binary subproblems. $\frac{4(4-1)}{2} = 6$ binary classifiers based on the pairwise combinations of the classes were built and evaluated. To determine the certainty of the predicted outputs, a post-processing step was performed that required a $\frac{2}{3}$ majority vote of the 6 classifiers in order to output a decision. By assessing the

certainty with a majority vote, the authors were able to determine which EEG classes the classifier had low confidence outputting. The classifier was able to achieve an 83% classification accuracy for the four grades [32].

Similarly, Ahmed et al. (2014) have built an automated grading system to classify four grades of EEG background activity using a combined Gaussian Mixture Model (GMM) supervector and SVM. They used the same dataset as Stevenson et al. for a direct comparison of results. First, the raw signals were preprocessed, and then features were extracted and decorrelated using Principal Component Analysis. A Universal Background Model (UBM) was created using a GMM, in which the means of the UBM were used to create a new GMM by using Maximum a Posteriori adaptation. The supervectors were created by concatenating the means into one, and were used as inputs to the SVM.

Six binary classifiers were created based on the pairwise combinations of the classes, and used for multi-class SVM classification with 2-fold cross-validation. The authors also determined the classifier certainty based on a $\frac{2}{3}$ majority vote of the 6 classifiers. This classification method yielded $> 90\%$ precision for grades 3 and 4 and $> 80\%$ for grades 1 and 2 [1].

Matić et al. (2014) used a tensor-based classification method to classify EEG recordings from full-term newborns who suffered from birth asphyxia into three grades of abnormalities, modified from Murray et al (2009). Hour-long epochs were graded without preselection by one neurophysiologist, and eight equidistantly-selected epochs of continuous EEG recordings were used from each newborn, resulting in a total of 272 1 hour epochs for the full dataset. The tensors were built by calculating the prevalence of three features that were computed from independent segments of the EEG channels. All three features correspond to the EEG amplitude since lower amplitude values are associated with more severe abnormalities. The first feature captured the peak-to-peak amplitude, the second feature captured the global, spatial amplitude, and the third feature captured the duration of segments that have the same amplitude class of low, medium, or high. The prevalence of the features with a particular feature distribution was calculated to obtain the 3D tensor structure. Dimensionality reduction was applied to the tensors using Tucker N-way decomposition, and feature selection was applied according to the Fisher information ranking score. Finally, classification was applied using the least-squares SVM (LS-SVM) algorithm with a one-versus-one coding scheme. The authors achieved an overall accuracy of 89% [22].

Matić et al. (2016) used a LS-SVM to classify EEG data from 53 neonates into levels of dynamic IBI (dIBI) that indicates both amplitude and duration of the detected IBI. The EEG recordings were split into three group to first optimize the detection of dIBI, then to optimize the post-processing test, and finally for validation. The first dataset used 157 hours of EEG segments from seven neonates with mild to moderate abnormalities,

the second dataset used 1 hour EEG segments from 38 neonates that contained milder abnormalities and were taken from the beginning of the recording. Neonates with mild and moderate EEG background discontinuities were used for validation since this group is typically the most difficult to discriminate. An neurophysiologist visually graded the EEG segments according to Murray et al.'s (2009) scoring system. The EEG segmentation was done with the same method as in Matic et al. (2012). Similarly, a low-amplitude temporal profile was also created to detect the dIBIs, and then a LS-SVM was applied to the detected dIBIs to further refine them into four groups: 1 ($3\text{ s} \leq \text{dIBI} < 5\text{ s}$), 2 ($5\text{ s} < \text{dIBI} < 10\text{ s}$), 3 ($10\text{ s} < \text{dIBI} < 5\text{ s}$), and 4 ($20\text{ s} \leq \text{dIBI} < 60\text{ s}$). The classifier achieved 95% true positive rate after refinement [20].

In this thesis, we classify full-term neonatal background EEG into discrete grades using support vector machine for multi-class and multi-label classification. Several differences between this thesis and some other works are: 1) We do not use start with the raw EEG signals, but rather the qEEG features as described in Appendix A 2) The EEG segments were not preselected for grading. The full duration of recording at 5-minute epochs for each patient was visually graded by two independent graders 3) The classification scheme used is based on amplitude voltage and IBIs, and we are grading discontinuity rather than just BSP patterns 4) We are attempting to determine how well the classifier performs if there is ambiguity between the graders' scores.

5. Machine learning background on classification

The simplest case in classification is the binary problem, in which the task is to classify an instance as either one of $K = 2$ non-overlapping classes. For example, one may be interested in automatically determining whether a patient should be diagnosed positive for a disease $(+1)$ or not (-1). Algorithms such as the support vector machine (SVM) and perceptron algorithm are designed to solve the binary cases [18].

The binary classifier can be extended to learning the multi-class problem. In multi-class classification, the learning problem is to classify the data into one of $K > 2$ non-overlapping classes. For instance, multi-class classification could distinguish between the three iris types in the Iris dataset.

Another subset of a classification problem is the multi-labeled problem. The binary and multi-class case mentioned above were described in the single-label case, in which each instance can be categorized into one of K classes. In the multi-labeled problem, some or all instances can be categorized into subsets of K classes. For example, a film could be classified as both the genre *comedy* and *horror*.

5.1 Support Vector Machines

Support vector machines (SVM) are well-studied for classification problems, particularly for text classification and in bioinformatics. Given an input set $\mathcal{X} \in \mathbb{R}^{n \times d}$ and an output set $\mathcal{Y} \in \{+1, -1\}^n$, SVM acts as a binary classifier and assigns x_i to y_i by mapping inputs into a higher dimensional space. A linear decision boundary of d-dimensions is learned so that it maximizes the margin between the support vectors.

The decision boundary is defined as $f(\mathbf{x}) : \mathbf{w}^T\mathbf{x} + b = 0$, where $\mathbf{w}^T\mathbf{x} = \sum_{i=1}^{n} w_i x_i$. Given a feature vector $\mathbf{x}$, the predicted output $\mathbf{y}$ is solved by $h(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$.

The margin forms the two half-spaces between two classes as defined by the decision boundary, and the support vectors are the data points that lie on the margin boundaries. A wider margin is preferred for less ambiguous predictions and allows for better generalization to new data [11].

Subsequently, the optimization problem is to maximize the margin with respect to finding the best parameters w and b that form the decision boundary. The objective function is

$$\text{minimize}_{w,b}\frac{1}{2}\|w\|^2 \tag{5.1}$$
$$\text{s.t. } y_i(w_i x_i) + b \geq 1, \text{ i=1...n}$$

The optimization problem is convex and of a quadratic form, allowing the classifier to avoid the local minima that other algorithms, such as neural networks, may face. There are two methods of setting up the optimization problem with Lagrange multipliers: the primal problem and the dual problem.

In an input space where there are n samples with d features, the dual function is preferable when $d \gg n$, and the primal function is preferable otherwise. This is because the dual method requires only the computation of inner products rather than minimizing over w, b subject to constraints. The dual and the primal are equivalent to an optimal convex solution if they satisfy the Karush-Kuhn-Tucker conditions.

SVM can solve the classification problem for non-separable inputs by using a soft margin, which allows margin violations with the addition of slack variables $\xi = (\xi_1, \xi_2, ..., \xi_n)$ and a penalty parameter C. Slack variables determine how many data points can violate the margin boundaries, and the penalty parameter determines the weight of the slack variables. With the addition of slack variables and a penalty parameter, the problem constraint is modified as:

$$y_i(w_i x_i) + b \geq M(1 - \xi_i) \tag{5.2}$$

$\forall i, \xi_i \geq 0, \sum_{i=1}^{n} \xi_i \leq$ constant. Equation 5.2 measures the overlap of the classes from the margin in relative distance. Bounding $\sum_{i=1}^{n} \xi_i$ to a constant c effectively bounds the total proportion of margin violations so that the maximum number of misclassified instances is c.

The optimization problem is subsequently reformulated as:

$$\text{minimize}_{w,b}\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \tag{5.3}$$
$$\text{s.t. } \xi \geq 0, y_i(w_i x_i) + b \geq 1 - \xi_i \forall i$$

As C increases, the margin width narrows, and there is a heavier penalty on assigning slack variables, causing the optimizing attempts to make unambiguous separation between the classes. Similarly, as C reduces towards 0, the margin width becomes wider, and there is less penalty on assigning the slack variables, which makes misclassifications more tolerable during optimization attempts. The hyperparameter **BoxConstraint** in

Matlab for SVM refers to the penalty parameter C. In the dual optimization problem, the Lagrange multipliers are constrained within the range $[0, C]$.

Non-linearity is supported through kernel mapping with a kernel that is chosen a priori, and the decision boundary equation is modified as $\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}) + b$. $\boldsymbol{\phi}$ is the kernel used in which $\mathbf{x}$ is mapped to. A kernel calculates the distance, or similarity, between two vectors by directly computing the inner products rather than computing the expanded representation. The kernel trick provides a computationally efficient way of projecting data into a higher-dimension [37].

A popular kernel to use is the Gaussian or Radial Basis Function (RBF) kernel: $K(\mathbf{x}, \mathbf{x}') = \exp(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2})$, in which we can substitute $\frac{1}{2\sigma^2}$ with the hyperparameter γ that determines the width of the kernel. The RBF kernel can be written as a dot product, $K(\mathbf{x}, \mathbf{x}') = \langle\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}')\rangle$, and is a measure of similarity where larger kernels correspond to closer distances between $\mathbf{x}$ and $\mathbf{x}'$. SVM predicts outputs with features that are similar to the learned features in the model under the assumption that similar instances would have similar outputs [37]. In Matlab, the **KernelScale** refers to γ.

To search for the optimal hyperparameters, a traditional approach is by using a cross-validation and grid search for C and for γ [11]. Since grid-search can be computationally expensive, it may be prudent to initially use a coarse grid in wider intervals to narrow down the search region before redefining smaller search intervals. Matlab provides a **bayesopt** hyperparameter optimization function that approximates the cross-validation rate with an internal k-fold cross-validation (default k = 5).

SVM on its own merely outputs the predicted classes. After the classifier outputs the weights of the support vectors, we can obtain the posterior probabilities of the output classes by using a modified sigmoid function. For each feature input x_i, we can evaluate the distance between x_i and the decision boundary, and then bound that metric between $[0, 1]$ for a probabilistic output [25]. If the classes are perfectly separated, Matlab transforms the scores to posterior probabilities by using the step function with some threshold c.

5.2 Class imbalance

In classification problems, class imbalance occurs when there are more instances in some classes that there are in others. The ratio at which the classes are imbalanced can be at 1:100 or larger. This creates a problem in classification when the classifier is unable to properly learn the rules for classifying the rarer classes, particularly in smaller datasets, so they end up being misclassified more often. Noisy data also makes it particularly difficult to discriminate the rarer cases.

Class imbalance is pervasive in anomaly detection, such as fraud detection and medical diagnosis, since the goal is to detect what is out of the norm. In situations where

it is valuable to classify rare cases, such as injury occurrences, it may be so that a higher identification rate for those classes is preferred.

5.3 Multi-class classification

The multi-class classification problem is a single-label problem where there is a finite set of $K > 2$ class labels, and each x_i is assigned to one $y_i \in \mathcal{Y} : \{1, ..., K\}^n$. It can be solved with a learning reduction approach, in which the multi-class problem can be decomposed into multiple binary subproblems. Algorithms that are efficient in solving binary problems, such as neural networks, k-Nearest Neighbor, and SVMs may then be applied to solve the subproblems. Their outputs will then be combined to obtain the final predictions [18].

Common methods of reduction include one-versus-all, and one-versus-one. Some advantages to these methods are that with the application of reductions theory, the reductions perform statistically well, are programmable, and can be used to solve large datasets [4].

The decomposition into binary classifiers is described using a code-matrix M. Each row of M indicates a different class, and each column of M indicates a set of labels that a different binary classifier will be learning (Figure 5.1). Each element could be of the values $\{-1, 0, 1\}$, where -1 indicates a negative class, $+1$ indicates a positive class, and 0 indicates that it does not participate in the classifier training. The K classes can be evaluated with $l = \lceil \log_2(K) \rceil$ binary classifiers, and there are $0.5(3^K + 1) - 2^K$ possible combinations of binary predictors in M [18]. In Matlab, **fitcecoc** allows for custom coding matrices.

5.3.1 One-versus-All

One method for reducing a multi-class problem into several binary problems is using the one-versus-all reduction, which creates K binary classifiers, $f_1, f_2, ..., f_K$, each classifier representing one positive class. The f_i model is trained so that those data in the i^{th} class are given the positive label, and the rest are given the negative label. M is represented as a $K \times K$ matrix, with its diagonal elements as $+1$, and the rest as -1. Each class is discriminated against the other $K - 1$ classes. Each new example is evaluated by the classifiers in which there are K decision functions, and the class receiving the highest score is the predicted output.

The number of errors ϵ that the binary classifiers can make in total is defined by the error rate of $(K - 1)\epsilon$; to reduce this error rate, it helps to randomly break ties and modify the weights of the classifier to favor positive outputs – thus potentially reducing

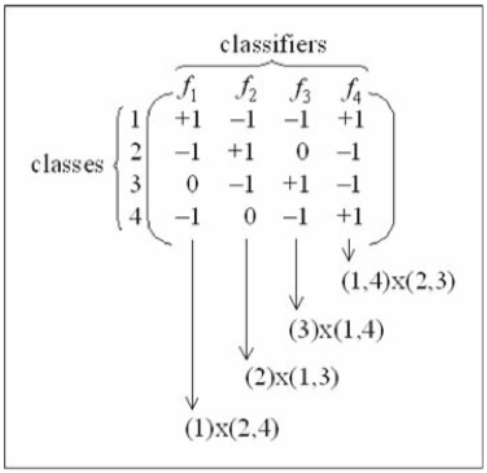

Figure 5.1: A four-class code-matrix (Lorena, 2008). The binary partition for each classifier is shown e.g., (1) × (2,4) indicates that classifier f_1 partitioned class 1 as the positive class and classes 2, 4 as the negative class. Class 3 did not participate in that classifier.

the error rate to approximately $\frac{K}{2}\epsilon$. However, the modifications to one-versus-all do not guarantee an improvement in cases where the binary problems are noisy [4].

5.3.2 One-versus-One

Another method for reducing a multi-class problem into a number of binary problems is using the one-versus-one reduction, which creates $\frac{K(K-1)}{2}$ binary classifiers for all pairwise combinations of the K classes. For all pairs of classes i, j and $i \neq j$, class i is discriminated against class j. M is represented as a $K \times \frac{K(K-1)}{2}$ matrix, where in each column, class i is $+1$, class j is labeled -1, and all other classes are labeled 0 since they are disregarded for that particular binary classifier.

The classifiers evaluate each new example and the predicted output is the class that receives a majority vote. Since not all predictors are involved with classifying the true class, they present themselves as noise in the final vote.

One-versus-one handles imbalanced datasets since it evaluates the classes in pairs, thus allowing the smaller classes to be discriminated against other individual classes, which potentially aids in separating the two classes. However, for datasets with few training instances, one-versus-one may risk overfitting the data.

5.3.3 Error-correcting output codes

In reductions theory, errors that are made during the reductions suggests that the original problem would manifest similar errors. Therefore, it is useful to measure these errors and reduce them. For multi-class problems in general, the error rates can be poten-

tially reduced by using error-correcting output codes (ECOC) for a distributed output representation. ECOC is robust to changes in training sample size and can provide reliable probability estimates for output classes [6]. In Matlab, it is used in the multi-class classification function **fitcecoc**.

The ECOC code-matrix can be described similarly to the code-matrix for decomposition problems. For each class, a codeword, or a unique binary string, of length n is assigned. The length n need not equal K, and a longer codeword is suggested to contribute to the robustness of the learning task. The codewords together form a binary matrix M of values $\{0, 1\}$, in which the rows correspond to each class. In the case of Figure 5.2, there are 10 classes and subsequently 10 codewords of length $n = 15$. Each $f_0 ... f_{14}$ represents a binary function to be learned for that respective column and evaluated to form an n-bit string [6].

Common evaluation of the codewords includes the Hamming distance, which evaluates how many bits apart two strings are (i.e., how many bits are different). For each new instance, a string is evaluated, and the Hamming distance is computed between the evaluated string to each of the ten codewords. The class corresponding to the closest codeword to the string according to the nearest Hamming distance will be the predicted class. For example, if the string '001110110000110' was evaluated, the nearest Hamming distance will be 4, corresponding to class 3 (Figure 5.2).

	Code Word															
Class	f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	Hamming distance
0	1	1	0	0	0	0	1	0	1	0	0	1	1	0	1	10
1	0	0	1	1	1	1	0	1	0	1	1	0	0	1	0	5
2	1	0	0	1	0	0	0	1	1	1	1	0	1	0	1	9
3	0	0	1	1	0	1	1	1	0	0	0	0	1	0	1	4
4	1	1	1	0	1	0	1	1	0	0	1	0	0	0	1	7
5	0	1	0	0	1	1	0	1	1	1	0	0	0	0	1	10
6	1	0	1	1	1	0	0	0	0	1	0	1	0	0	1	8
7	0	0	0	1	1	1	1	0	1	0	1	1	0	0	1	9
8	1	1	0	1	0	1	1	0	0	1	0	0	0	1	1	9
9	0	1	1	1	0	0	0	0	1	0	1	0	0	1	1	8
String	0	0	1	1	1	0	1	1	0	0	0	0	1	1	0	

Figure 5.2: A 15-bit ECOC for $k = 10$ (Modified from Dietterich, 1994).

For ECOC to effectively reduce the error, the codewords should be well-separated, and the individual errors from the binary classifiers must be uncorrelated. Otherwise, identical or columns that are complements of each other will produce identical errors. Dietterich et al. (2004) suggests four techniques based on the number of classes to meet these criteria, though the authors did not justify how the number of classes were stipu-

lated. For problems where $K \leq 7$, they suggest using an exhaustive code, which would generate $2^{(K-1)} - 1$ binary classifiers with Hamming distance being $2^{(K-2)}$ [6].

Since ECOC is a more complex decomposition into binary subproblems, these problems are subsequently challenging to learn and require longer training time, despite having good error-reducing properties. By iteratively experimenting with adding and removing ECOC codewords and choosing promising ones according to its error-correcting capabilities, this introduces an additional function that can evaluate a codeword's quality. The length of the codeword can also be systematically determined by evaluating the number of errors that different lengths of ECOCs can correct [18].

5.3.4 SVM as a base classifier

Lorena et. al. (2008) suggests using SVM for multi-class problems since they generally perform well, though they have high computational complexity and subsequently long training time. SVM can be extended for multi-class problems by training multiple binary classifiers with the approaches mentioned in section 5.3. Another approach for multi-class SVM is to directly compute the optimization problem for all classes, creating variables that associate with the number of classes. A similarity that both approaches share is that the problem space is expanded. However, training multiple binary classifiers still takes less computational resources than solving a larger optimizing problem [12].

The one-versus-one approach is suggested to perform better than one-against-all in a comparative study using ten multi-class datasets from the UCI repository [12]. This may be due to one-vs-one discriminating between fewer classes each time during optimization. This approach does require higher complexity of K^2 order due to the increased combination of classifiers from the pairwise combinations to evaluate. However, since each classifier is of the binary case, it does not require too much time to compute [18]. Particularly for SVMs, the computational time depends more on the number of unique support vectors to evaluate than the number of decision functions to evaluate. One-versus-all uses more support vectors than one-versus-one does, so the training time for one-versus-all is significantly higher than one-versus-one. While there are computational savings that could be implemented by avoiding recalculating the same kernel matrix several times, the training time is still expected to be higher for one-vs-all [12].

SVM also lessens the effects of an imbalanced class distribution. While many algorithms are inadequate in learning with class imbalance, SVM is often less sensitive to the imbalanced distribution since decision boundaries only depend on several support vectors, regardless of class size. However, in extremely imbalanced cases, the support vector ratio between the prevalent classes and rarer classes increases. SVM subsequently favors predicting the prevalent classes in order to maximize the margin. In the multi-class imbal-

anced case, using the one-versus-all method may worsen the performance of the smaller class since the classifier now has a massively imbalanced problem in the binary case [33].

5.4 Multi-label classification

Multi-class problems are often single-label problems, in which each instance is associated with one class label. In contrast, multi-label problems, some instances are associated with more than one class labels in a finite set of K class labels. This is common in text classification of movie genres where a film could be categorized as 'horror' and 'comedy' [35].

Multi-label datasets can be described with the statistics, label cardinality and label density. Label cardinality refers to the average number of labels per instance, and label density normalizes label cardinality by the total number of class labels K [41].

There are two main methods to solving multi-label problems – the problem transformation methods, and algorithm adaption methods. The former method entails transforming the problem into a single-class algorithm prior to evaluation, and the latter method entails directly extending learning algorithms to classify multi-labels.

One naive method for transforming the problem into a single-label problem is to randomly discard the labels so that there is only a single label for each feature input. If the labels are continuous, one could take the mean or median between the two classes. However, the former method also discards a lot of information from the original dataset, and the latter does not extend well to categorical labels [35].

Several more robust problem transformation methods include the binary relevance method (BR), binary pairwise classification method (PW), label combination/label power-set method (LC), and copy-weight transformation method (CW) [30].

5.4.1 Binary relevance method

The most common approach to transforming the multi-label problems is the binary relevance method (BR), which is to learn K binary classifiers for each of the different labels, similar to the single-label multi-class classification problem (Table 5.1a) [8]. However, BR assumes label independence, so if there are label dependencies in the data, BR cannot directly capture label correlations and may not perform optimally [30]. Moreover, it may not perform well if the label density is low [41].

5.4.2 Binary pairwise classification method

The binary pairwise classification method (PW) is to learn $\frac{K(K-1)}{2}$ binary classifiers for each *pairs* of the labels, resulting in pairwise outputs of labels (Table 5.1b). PW is prevalent in use with ranking schemes, in which the labels can be ranked according to its relevance to the instance. A disadvantage of this method is the increased computational power as its complexity is quadratic, which prevents PW as being ideal for practical use [35].

5.4.3 Label combination/label power-set method

The label combination/label power-set method (LC) combines the multi-labels into one atomic label, effectively creating new single-labels (Table 5.1c). For example, a label $y_i = \{0, 1\}$ would be transformed into $y_i =$'01'. This has the potential to increase the number of classes to an exponential complexity, depending on the number of subsets of label groupings, and makes this an unideal method to use when there are many labels involved. Moreover, LC tends to overfit during training since it only observes the subsets of labels in the training instances.

To mitigate these problems, a related method, RAkELd, was developed. It randomly groups the labels into smaller subsets before combining into a single label, thus reducing the computational load as well as preventing a skewed class distribution.

5.4.4 Copy-weight transformation method

Copy-weight transformation is to duplicate the feature inputs with multiple labels and to create a weight vector with weights of $\frac{1}{k_i}$, where k_i corresponds to the number of labels belonging to the i^{th} instance (Table 5.1d).

A disadvantage of this method is that when using SVM to classify the instances, it is unclear where the decision boundary should lie when the features overlap but appear to belong in different classes, as it makes the classes inseparable. Moreover, as the set of labels grow, so does the data set. Similarly to the binary approach, this transformation does not take into account label dependencies.

5.4.5 Label dependencies

Regardless of which method is used to transform multi-label into a single-label problem, the consensus is that the set of labels are not necessarily independent. This may lead to ambiguity in the feature space due to feature overlap corresponding to multiple classes. The dependencies between the labels may be an informative feature to use for training

a) Binary Relevance

instance	$y = 1$	$\neg y = 1$
x_1	0	1
x_2	1	0
x_3	1	0
x_4	0	1

instance	$y = 2$	$\neg y = 2$
x_1	1	0
x_2	0	1
x_3	1	0
x_4	1	0

instance	$y = 3$	$\neg y = 3$
x_1	1	0
x_2	0	1
x_3	1	0
x_4	0	1

instance	$y = 4$	$\neg y = 4$
x_1	0	1
x_2	0	1
x_3	0	1
x_4	1	0

b) Binary Pairwise

instance	label
x_1	$\neg y = 1, y = 2$
x_2	$y = 1, \neg y = 2$
x_4	$\neg y = 1, y = 2$

instance	label
x_1	$\neg y = 1, y = 3$
x_2	$y = 1, \neg y = 3$
x_4	$\neg y = 1, \neg y = 3$

instance	label
x_2	$y = 1, \neg y = 4$
x_3	$y = 1, \neg y = 4$
x_4	$\neg y = 1, y = 4$

instance	label
x_4	$y = 2, \neg y = 3$

instance	label
x_1	$y = 2, \neg y = 4$
x_3	$y = 2, \neg y = 4$

instance	label
x_3	$y = 3, \neg y = 4$
x_4	$\neg y = 3, y = 4$

c) Label Transformation

instance	label
x_1	$y = 23$
x_2	$y = 1$
x_3	$y = 123$
x_4	$y = 24$

d) Copy Weight Transformation Method

instance	label	weight
x_{1a}	$y = 2$	0.5
x_{1b}	$y = 3$	0.5
x_2	$y = 1$	1
x_{3a}	$y = 1$	0.33
x_{3b}	$y = 2$	0.33
x_{3c}	$y = 3$	0.33
x_{4a}	$y = 2$	0.5
x_{4b}	$y = 4$	0.5

Table 5.1: Problem transformation methods (Modified from Sorower, 2010). Suppose there is the training set $(x_1, y = \{2, 3\}), (x_2, y = 1), (x_3, y = \{1, 2, 3\}), (x_4, y = \{2, 3\})$. **a** illustrates the binary relevance transformation, **b** illustrates the binary pairwise transformation, **c** illustrates the label transformation, and **d** illustrates the copy weight transformation.

the classifier. A method to capture the co-occurrence patterns between classes is to define additional parameters over pairs of labels and features [8].

5.5 Classifier performance evaluation metrics

A classifier's outputs can be organized into a confusion matrix to evaluate its performance. A confusion matrix (Figure 5.3) can be used to calculate the overall accuracy, which is a

single score that considers the number of true positives over the total number of samples.

Data class	Classified as *pos*	Classified as *neg*
pos	true positive (*tp*)	false negative (*fn*)
neg	false positive (*fp*)	true negative (*tn*)

$$\begin{bmatrix} tp & fn \\ fp & tn \end{bmatrix}$$

Figure 5.3: A confusion matrix for binary classification (Sokolova, 2009).

However, binary, single-label evaluation metrics do not present the classifier performance for individual categories nor take into account of instances associated with several labels. Several evaluation metrics address these differences for multi-class and multi-label classifiers (Table 5.2) [18].

For multi-class classification, measures such as precision, recall, and the F-measure can be used to evaluate each class label independently. Average accuracy describes the average accuracy of individual class labels. Two variants of averaging across the class labels are micro-averaging and the macro-averaging.

Taking the macro-average means to average the K measures calculated for K individual labels, while taking the micro-average means to average a single measure over n instances. The micro-average favors larger classes since they would contribute more when averaging over smaller classes. In contrast, the macro-average assumes that all classes contribute to the classifier equally, which may make it better suited to evaluate an imbalanced dataset [29].

Two main methods to evaluate multi-label problems are example-based metrics and label-based metrics. Example-based metrics include taking the mean value of accuracy, precision, recall, F-measure, Exact match ratio, or Hamming Loss across the test set. Label-based methods include evaluating by individual class labels, and then taking the micro/macro-average of accuracy, precision, recall, and F-measure [41].

Multi-label evaluation metrics can also be organized as partial or exact class label matching. Partial class label matching considers that the classifier may not always be able to acquire an exact distribution of each class from the training instances, and takes into account each element in a label subset equally. In contrast, exact class label matching is a much stricter metric since all elements in a label subset must agree with the true label to be considered a true positive. Example metrics of partial and exact class label matching is the Hamming Loss and Exact Match Ratio [29].

A measure is invariant if its value does not change even if a set of values in the confusion matrix changes. If the goal of a classifier is to find true positives, precision and recall are beneficial to use since they describe a classifier's ability to detect positive classes, and display invariance for true negatives. In contrast, non-invariant measures to true negative classes describe a classifier's ability to detect negative classes [29].

Multi-class Measure	**Formula**
Average Accuracy	$\frac{1}{K}\sum_{i=1}^{K}\frac{tp_i+tn_i}{tp_i+fn_i+fp_i+tn_i}$
Precision_μ	$\frac{\sum_{i=1}^{K} tp_i}{\sum_{i=1}^{K} tp_i+fp_i}$
Recall_μ	$\frac{\sum_{i=1}^{K} tp_i}{\sum_{i=1}^{K} tp_i+fn_i}$
F1-score_μ	$2\times\frac{\text{Precision}_\mu\times\text{Recall}_\mu}{\text{Precision}_\mu+\text{Recall}_\mu}$
Precision_M	$\frac{\sum_{i=1}^{K}\frac{tp_i}{tp_i+fp_i}}{K}$
Recall_M	$\frac{\sum_{i=1}^{K}\frac{tp_i}{tp_i+fn_i}}{K}$
F1-score_M	$2\times\frac{\text{Precision}_M\times\text{Recall}_M}{\text{Precision}_M+\text{Recall}_M}$
Multi-label Measure	**Formula**
Exact Match Ratio / Subset Loss	$\frac{1}{n}\sum_{i=1}^{n} I(\hat{y}_i = y_i)$
Hamming Loss	$\frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} I(\hat{y_{ij}} \neq y_{ij})$

Table 5.2: Summary of evaluation metrics for multi-class and multi-label classification (Modified from Sokolova, 2009). μ represents micro-averaging, M represents macro-averaging. K is the total number of class labels, and tp, fp, tn, fn denotes the true/false positive/negative counts. I is the indicator function. y_i denotes the true class for instance i, and $\hat{y}_i$ denotes the predicted class. k is the total number of labels in a subset.

6. Methods

The main goals of this thesis are to create a classifier that could distinguish between levels of continuity in EEG patterns of full-term neonates and describe the cerebral development of the newborns with a visualization that could be used by neonatologists that may or may not be experts in EEG.

To analyze the full duration of EEG recording, we would have to reconcile the manual grading of the epochs by the independent graders. We begin by creating two classifiers that use the two graders' individual labels as a baseline reference. Then, we select epochs that were graded in consensus to train another classifier and use it to evaluate the epochs in disagreement. If the features of the ambiguous class labels correlate well with the features of epochs graded in consensus, then the classifier should improve over the reference classifiers.

Finally, we directly learn the full-duration recording by using a multi-label approach to explore how well the classifier can reconcile the scores in disagreement. We considered that both graders captured the EEG transitions satisfactorily to a trained physician's visual observation, despite having 5678 epochs (40% of the total recording) in disagreement. Therefore, we assumed that the two interpretations of the EEG patterns inherently contain latent information that the classifier can exploit. Subsequently, we can treat this as a multi-label classification problem by decomposing the multi-labels into single labels. In other words, where there is a disagreement in the EEG grade between the two graders, we consider both labels as the true class labels so that some instances can be classified into a maximum of two grades.

6.1 Data acquisition

The datasets used in this thesis were acquired from the BABA Center. The original EEG datasets were provided to the BABA Center by a collaboration between Professors Emily Tam and Cecil Hahn (Toronto) and Professor Sampsa Vanhatalo (Helsinki). The data was obtained from a clinical trial, Neurological Outcome of Glucose in Neonatal encephalopathy (NOGIN), that was conducted at the The Hospital for Sick Children (SickKids) in Toronto, Canada. The study was partly aimed at studying long-term glucose

monitoring during NICU treatment after birth asphyxia.

Continuous multichannel EEG recordings were taken from 31 full-term neonates using Stellate Harmonie or Xltek Brain Monitor ICU video-EEG systems (Natus Neurology, Oakville, Ontario, Canada) with a 12–14 channel bipolar montage. The EEG recording continued over 12–48 hour periods, capturing sleep-wake activity. Visual review and background scoring were performed by two neurologists, Drs. Elana Pinchevsky (Toronto) and Viviana Marchi (Pisa/Helsinki), blindly.

Data preprocessing was performed by Ms. Minna Kauppila, while feature calculation was performed by Ms. Karoliina Tapani. The 75 qEEG features are listed in Appendix A.

6.1.1 EEG pattern labels

Two neurologists, E.P. and V.M, independently annotated the EEG patterns by referring to the Toronto epoch timestamps and using the scoring system by The Hospital for Sick Children at the University of Toronto as described in Table 6.1. The Toronto timestamps list the patient id, epoch date, and epoch time by increments of 5 minutes.

Epochs of 5-minute intervals were used for the visual interpretation of the EEG data and can be classified into one of eight scores (Table 6.1). Some epochs were not graded if there were excessive movement artifact or seizures.

Score	Description
0	**continuous** (recovered from inactivity / poor activity)
1	**trace alternant**: IBI voltage $\geq 25\mu$V with IBI duration ≤ 6 seconds
2	**trace alternant**: IBI voltage $\geq 25\mu$V with IBI duration > 6 seconds
3	**trace discontinuous**: IBI voltage $< 25\mu$V
4	**depressed and undifferentiated**: persistent low-voltage background activity w/ amplitude between 5μV and 15μV and w/o normal features
5	**burst suppression**: IBI amplitude $< 5\mu$V
6	**very low voltage**: amplitude $< 5\mu$V or with no discernible cerebral activity
999	**unable to assess background** (due to seizure or artifact)

Table 6.1: cEEG scoring guide (Courtesy of BABA Center). IBI: inter-burst interval. This scoring system was created for use in the NOGIN clinical study, and it describes the EEG abnormality that may be seen after birth asphyxia by classifying the background patterns into seven grades. The EEG behavior states here are described with amplitude ranges and burst patterns with defined duration. SWC is not used by this system since the patterns are graded every five minutes, as opposed to other clinical set ups that may grade the EEG by the hour. Moreover, this scoring system describes seven levels of EEG pattern abnormality, which is more detailed than other scoring schemes.

Twenty-seven individual subjects' EEG data were annotated instead of 31 since four patient files were corrupted. We received two sets of EEG pattern grades, and we then

aligned the two graders' labels according to the epoch timestamps. Epochs with score 999 were indistinguishable signals, so we removed them from the dataset. This subsequently reduced the total number of class labels from eight to seven.

After aligning the epochs and removing missing or undistinguishable instances, the dataset contained a combined length of 14103 epochs (approx. 1175 hours) of recording. The average length of the recordings per patient is 44 hours (range 9–100 hours). Of those epochs, 5678 epochs (approx. 473 hours in total) were in disagreement (for the average patient, 210 epochs or 17.5 hours). Cohen's unweighted kappa statistic is κ=0.45, which suggests moderate agreement of EEG grades, and subsequently fairly good reproducibility. The raw proportion of agreement is 0.60.

Since we will refer to the graders a considerable amount for the remainder of this thesis, E.P. and V.M. will be referred to as 'Grader 1' and 'Grader 2' respectively for clarity. Table 6.2 describes the distribution of the class labels. There is a class imbalance where the EEG pattern labels are skewed towards grade 0 (Graders 1 and 2) and grade 3 (Grader 1), compared to the significantly fewer instances of grade 4, 5, 6 (Graders 1 and 2).

	Grader 1		**Grader 2**	
Grade	**Count**	**Percent**	**Count**	**Percent**
0	7350	52.12	5756	40.81
1	805	5.71	2196	15.57
2	1113	7.89	1838	13.03
3	3346	23.73	1822	12.92
4	254	1.80	742	5.26
5	762	5.40	1199	8.50
6	473	3.35	550	3.90

Table 6.2: Tabulated summary of the graders' labels. Both graders classify a larger proportion of EEG grades as 0 or 3.

As seen in Figure 6.3, Grader 2 tended to classify more patterns as moderate/severe than Grader 1 (i.e., disagreement between grades 5 and 3). There are no disagreements in which Grader 2 classified the EEG trace as normal/mild and Grader 1 classified as moderate/severe.

		Grader 1 vs Grader 2 scores						
	0	**5283**	1412	533	111	8	3	0
	1	157	**329**	255	64	0	0	0
	2	178	253	**461**	221	0	0	0
Grader 1	**3**	138	202	589	**1374**	262	781	0
	4	0	0	0	0	**164**	1	89
	5	0	0	0	52	251	**406**	53
	6	0	0	0	0	57	8	**408**
		0	**1**	**2**	**3**	**4**	**5**	**6**
		Grader 2						

Table 6.3: EEG grades assessed by the two graders. The bolded diagonal entries describe grades scored in consensus, and the unbolded entries describe the grades scored in disagreement.

6.2 Pre-processing the features

We assume HIE leads to diffuse cerebral dysfunction, which allows us to take the median across the channels for each feature. Then, we standardized the data with z-score standardization to zero mean and unit variance, which bounds all features within $[0, 1]$. Scaling is an essential pre-processing step to take prior to using SVM to prevent any features from dominating the rest due to an extreme scale.

6.2.1 Feature selection

Due to the high number of extracted features from the EEG data, there is the risk of increased computational time as well as overfitting. To mitigate this, we also experiment with using stepwise regression to obtain the best subset of features to use for training. In stepwise regression, features are systematically added and removed in a regression model depending on whether its p-value meets a tolerance level. In Matlab, the function for stepwise regression algorithm is **stepwisefit**.

6.3 Datasets

We derived four datasets to use in creating the different classifiers. All of the four dataset uses the same feature set, but with different variations of selected epochs and class labels.

Dataset 1 and 2 contain all epochs (those graded in consensus and disagreement) labeled by Grader 1 and Grader 2 respectively. Both consists of 14103 combined epochs. We will evaluate these two independently as reference models.

Dataset 3 contains only epochs that were graded in consensus, which consists of 8425 combined epochs. This allows us to interpret the classifier results without any ambiguity from the truth of the labels. We later use this model to evaluate how well it performs upon seeing inherently ambiguous instances (epochs in which the graders disagreed on) by predicting Grader 1 and Grader 2's labels independently.

To create **dataset 4**, we expanded the feature space by duplicating the instance with class labels that the graders disagreed on, corresponding to the copy-weight problem transformation approach in multi-label learning. In addition, we created a corresponding weight vector where each class label in consensus was given a weight of 1, and each class label in disagreement that was duplicated was given a weight of 0.5. The dataset subsequently grew to a combined length of 19781 epochs. The label cardinality is 1.4, and label density is 0.2.

For the remainder of this thesis, we will refer to dataset 1 as *grader 1*, dataset 2 as *grader 2*, dataset 3 as *consensus*, and dataset 4 as *CW*. Figure 6.1 summarizes the different models trained by the datasets.

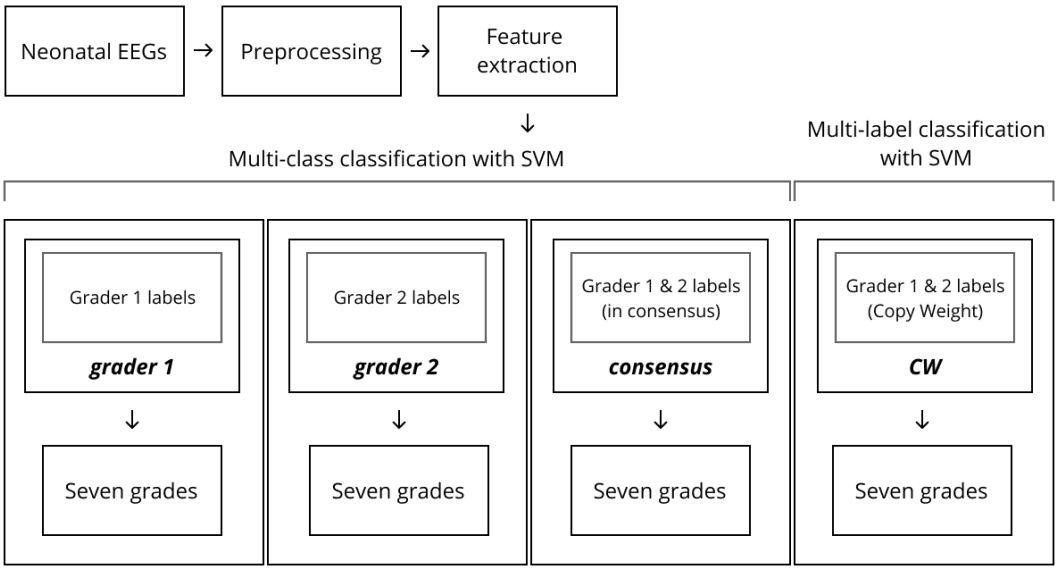

Figure 6.1: Summary of different classification models built with the datasets *grader 1*, *grader 2*, *consensus*, and *CW*.

6.4 Experimental setup

For all four datasets, we randomly shuffled the data by patients to produce a random 70/30 split for training and testing respectively. Shuffling the patient order allows us to preserve the epoch orders by patients since the features for each epoch per patient have

time dependencies. More importantly, this prevents test set contamination that may occur if we permute by correlated epochs.

We then used the same procedure on *grader 1*, *grader 2*, and *consensus* in Matlab by using **fitcecoc** to train a multi-class ECOC model that uses SVM as a base classifier with a Gaussian kernel on the training set. We use a one-vs-one coding system to create 21 binary learners. We chose a one-vs-one approach to avoid exacerbating the problem of class imbalance.

We also used **fitcecoc** to train the multi-label classifier using CW due to its multi-class properties, but we gave the weight vector as an additional input to the classifier. The weight vector is then normalized so that its sum equals the prior probability of the classes.

SVM with a Gaussian kernel can be optimized with the hyperparameters, **BoxConstraint** (C) and **KernelScale** (γ), which corresponds to the regularization parameter to penalize slack variables, and the width of the kernel respectively. To find the optimal hyperparameters, we used grid search from the intervals $[2^{-5}, ..., 2^{15}]$ for choosing C, and $[2^{-15}, ..., 2^{3}]$ for choosing γ [11]. We decided to use grid search since there are only two parameters to search for, thus limiting the required computational time, and grid search also allows for parallelization since the hyperparameter pairs are independent. Although Matlab's hyperparameter optimization has an option to specify grid search as its optimizer, users can only define the number of values that the optimizer will randomly evaluate along some grid, rather than a specified search range of values. As such, we implemented an exhaustive search instead. In total, we searched $21 \times 19 = 399$ parameter combinations, using Matlab's Parallel Computing Toolbox to run the computations in parallel.

For each parameter pair, we used 5-fold cross-validation, using 5 patients as validation each time, and selecting the parameter pair with the highest average cross-validation accuracy. Using the entire training set, we retrained the model with the selected parameter pair and then evaluated the classifier on the test set. To see how well the model generalizes to new data, we used leave-one-out cross-validation (LOO-CV), leaving out one patient for validation each time, to evaluate the full datasets and report the average accuracy. We also transformed the classification scores into posterior probabilities for visualization in Section 7.6. All experiments were performed on either a 2-processor 2.7 GHz i5 machine with 8GB ram, running Mac OS, or an ukko2 cluster with 48 cores provided by the University of Helsinki.

In creating the visualizations to illustrate the outputs of the classifier, we considered that physicians prefer to see a confidence score for the predicted classes, and that they are interested in the development of the newborn. EEG signals naturally transition between grades throughout the normal sleep-wake cycle. Since the EEG is scored over five-minute epochs, the recorded grades will reflect the cycles as noise, rather than the overall state

of the EEG. By smoothing the grades, we can filter out natural transitions of the sleep-wake cycle that do not contribute much to determining the neurological development of the newborn. Thus for part of our visualizations, we observe the overall development of the trend by taking the moving average of the labels over 4 and 12 five-minute epochs, smoothing the EEG grades by 20 and 60 minutes.

6.5 Evaluating classifier performance

To evaluate the classifiers, we used raw accuracy, average accuracy, and several macro-averaging metrics of precision, recall, and F1 score. We use macro-averaging measures due to the imbalanced dataset. We also use Hamming Loss to evaluate the multi-label classifier, and not Exact-Match Ratio since we are satisfied if the classifier can predict at least one of the graders' score. We evaluate the classifiers using the test sets, and also determine how well the classifiers can generalize to new data by using LOO-CV on the full datasets.

Finally, we evaluated the test set of *grader 1* and *grader 2* with the *consensus* model, as we hope the classifier had learned the features of the grades in consensus. This allowed us to evaluate how well the classifier extends to ambiguous labels (i.e., EEG pattern grades in disagreement between the two graders).

We expect *consensus* to perform the best since it is likely to have distinctive features that reflect the grades in agreement. We also hypothesize that this model can learn generalizable relationships to predict ambiguous instances. Therefore, we will evaluate the test set of *grader 1* and *grader 2* with the *consensus* model. If the *consensus* model is able to learn the features of the epochs that the graders agreed on, then it could act as a noise filter for the features that make the epochs ambiguous to grade. If so, then the subsequent performance on the test sets would improve compared to the baseline models – those trained with the graders' respective labels and which contains features of epochs in disagreement. If the *consensus* model cannot filter out the noise, we expect similar performance to the baseline models.

It is important to note that the pattern grades have a slight ordinal relationship in that lower grades are associated with better outcome, and worsens as the grades increase. However, this relationship is not linear, and in reality, physicians would consider grade 0 as the healthiest, $[1, 2, 3]$ as still healthy, and $[4, 5, 6]$ as abnormal.

To closer reflect how physicians interpret the scores, as well as to manage the class imbalance that is skewed towards grades 0 and 3, we combined the seven labels into three classes: 0, $[1, 2, 3]$, and $[4, 5, 6]$. This reveals which models misclassifies more severely into classes that are farther away, since misclassifications that are close to the true class (i.e., falls within the groups) can be accepted.

7. Results

The highest performing model was with the *consensus* dataset at 82% overall accuracy (95% average accuracy) on the test set and 79% overall accuracy with LOO-CV. Table 7.1a summarizes the results between the different models, and Table 7.1b displays the confusion matrix of average outputs from LOO-CV over the full datasets.

7.1 *grader 1* and *grader 2*

The datasets *grader 1* and *grader 2* were independently trained as a baseline reference to compare against the *consensus* model. The classifiers achieved 71% and 57% for *grader 1* and *grader 2* respectively, 92% and 88% average accuracy respectively, and 68% and 55% LOO-CV accuracy on the full dataset respectively.

The test set that was used to evaluate the trained model did not contain true class label 4 for *grader 1*, which motivated the evaluation of the full dataset with LOO-CV (Table 7.1b). *grader 1* had high recall, or true positive rate, for grades 0 and 3 (Table 7.5), while *grader 2* had high recall for only grade 0.

Although the classifier was able to identify less true class labels for other classes, the misclassifications tend to occur at neighboring classes for both datasets (Tables 7.5, 7.6). When using Grader 2's class labels, the grade 5 was often misclassified as 3 or 4 grade, and the grade 6 was often misclassified as grade 4. Grader 2 had labeled more instances as grade 6, and the *grader 2* model subsequently was able to identify more instances as so (54% recall).

7.2 *consensus*

The *consensus* model out-performed the individual reference models with 82% overall accuracy (95% average accuracy) for the test set, and had a 79% LOO-CV for the full dataset.

To evaluate how well the *consensus* model can predict class labels for features that were graded in disagreement, we evaluated the test set of *grader 1* and *grader 2* (both

	grader 1		grader 2		consensus		CW	
	C	γ	C	γ	C	γ	C	γ
	8.00	8.00	0.25	8.00	8.00	8.00	8.00	8.00
5-Fold CV avg raw accuracy (train set)	0.70		0.57		0.78		0.67	
Raw accuracy (test set)	0.71		0.57		0.82		0.67	
Accuracy w/ grouped class labels (test set)	0.78		0.69		0.87		0.74	
Avg accuracy (test set)	0.92		0.88		0.95		0.91	
Precision$_M$ (test set)	0.44		0.42		0.50		0.43	
Recall$_M$ (test set)	0.32		0.38		0.36		0.31	
F1-Score$_M$ (test set)	0.37		0.40		0.42		0.36	
Hamming Loss (test set)	-		-		-		0.17	
LOO-CV avg raw accuracy (full)	0.68		0.55		0.79		0.65	

(a) Summary of evaluation metrics on the models, trained with all features. M denotes macro-averaging.

consensus (AVG ACC 79%)

True Class \ Predicted Class	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	**190**	1	4	3	1	0	0
1	10	**2**	2	1	0	0	0
2	8	1	**7**	3	0	0	0
3	4	1	5	**40**	1	4	0
4	5	0	0	1	**1**	0	2
5	1	0	0	13	0	**2**	1
6	1	0	0	1	4	1	**10**

CW (AVG ACC 65%)

True Class \ Predicted Class	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	**322**	2	10	16	1	1	0
1	39	**3**	4	4	0	0	0
2	40	2	**13**	12	0	0	0
3	34	2	27	**123**	1	9	4
4	4	0	0	2	**1**	1	9
5	2	0	0	35	1	**5**	2
6	1	0	0	4	9	1	**7**

grader 1 (AVG ACC 68%)

True Class \ Predicted Class	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	**258**	1	6	9	1	1	0
1	25	**2**	2	2	0	0	0
2	25	1	**9**	7	0	0	0
3	20	1	16	**83**	1	4	3
4	4	0	0	2	**1**	1	5
5	1	0	0	24	1	**3**	1
6	1	0	0	4	7	2	**7**

grader 2 (AVG ACC 55%)

True Class \ Predicted Class	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	**200**	5	7	2	1	1	0
1	62	**7**	13	1	0	0	0
2	33	5	**25**	6	1	1	0
3	12	1	19	**26**	1	12	0
4	8	0	1	1	**7**	11	3
5	3	0	1	25	10	**9**	1
6	2	0	0	0	8	1	**12**

(b) Average confusion matrix from LOO-CV over full dataset. The misclassifications tend to skew towards 0 when the true class is 1, 2, or 3.

which contains epochs in disagreement) with the *consensus* model. *grader 1* achieved 70% overall accuracy, and *grader 2* achieved 59% overall accuracy. Compared to the baseline

grader 1

True Class \ Predicted Class	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	**1923**	38	129	236	13	0	2
1	127	**30**	40	14	0	0	0
2	151	18	**75**	136	1	0	0
3	9	1	7	**448**	7	49	0
4	0	0	0	0	**0**	0	0
5	3	0	0	12	21	**36**	43
6	4	0	0	1	73	12	**140**

grader 2

True Class \ Predicted Class	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	**1601**	12	59	220	6	0	2
1	400	**35**	33	23	4	0	0
2	159	33	**85**	122	2	0	0
3	49	7	73	**351**	8	49	0
4	1	0	0	108	**16**	23	11
5	0	0	1	22	1	**0**	1
6	7	0	0	1	78	25	**171**

Table 7.2: Confusion matrix of *grader 1* and *grader 2* test set evaluated with *consensus* model.

performances in section 7.1, there was no improvement in classifying the features graded in disagreement.

7.3 *CW*

The copy-weight dataset was the expanded dataset that included grades in disagreement and used both graders' grades as class labels. The overall accuracy for the test set was 67% (91% average accuracy, 65% LOO-CV for the full dataset). Similarly to *consensus* and *grader 1*, there was high recall in grades 0 and 3. The Hamming Loss was 0.17, which indicated that the proportion of incorrect predictions was comparable to other methods.

7.4 Selected features

During training, we used 5-fold cross-validation to search for 399 hyperparameter pairs for the SVM model. During the nested loop for each fold, we used **stepwisefit** to repeatedly select features to generate the optimal feature set for each model, thus searching on average 1967 combinations of feature combinations for each model. Table 7.3 summarizes the commonly selected features.

The overall accuracies remained the same for the models that used a reduced set of features when compared to those that used the full set of features (Refer to 7.5).

7.5 Combined grade evaluation

By combining the grades into three groups: 0, $[1, 2, 3]$, and $[4, 5, 6]$, we effectively allowed neighboring classes to be considered as correct.

As expected, the accuracy after grouping by class labels increased for all models, notably for *grader 2*. A significant number of $[1, 2, 3]$ was misclassified as 0, but rarely

Most frequently selected features	Number of times selected on average	Most frequently selected features (continued)	Number of times selected on average (continued)
amp_variance	1967	band1_3norm	1717
band10_12norm	1967	AR4	1714
burst_nro	1967	H_shannon	1707
Line_length	1935	m_med	1679
fractal_dim	1912	mdfa	1548
band5_7norm	1905	AR1	1470
Cov_IA_IF	1899	freq_variance	1444
band2_4norm	1891	rEEG_5	1418
band8_10norm	1851	H_spectral	1395
peak_freq	1821	freq_kurt	1361
dfa	1781	freq_mean	1343
burst_duration	1767	fisher	1335
wavelet_energy	1741	Zero_crossings	1167

Table 7.3: The commonly selected features across all 4 models, on average over 1967 combinations of selected features.

was grade 0 misclassified as $[4, 5, 6]$, and vice versa.

consensus

True class	0	1, 2, 3	4, 5, 6
0	**1636**	63	27
1, 2, 3	141	**327**	83
4, 5, 6	3	8	**227**

Predicted class

CW

True class	0	1, 2, 3	4, 5, 6
0	**2678**	234	44
1, 2, 3	772	**848**	55
4, 5, 6	3	220	**229**

Predicted class

grader 1

True class	0	1, 2, 3	4, 5, 6
0	**2158**	152	31
1, 2, 3	474	**607**	32
4, 5, 6	3	139	**203**

Predicted class

grader 2

True class	0	1, 2, 3	4, 5, 6
0	**1764**	114	22
1, 2, 3	792	**398**	243
4, 5, 6	5	15	**446**

Predicted class

Table 7.4: Confusion matrices for grouped evaluation

consensus (ACC 82%)

True Class	0	1	2	3	4	5	6
0	**1636**	1	3	59	24	0	3
1	75	**2**	2	0	0	0	0
2	66	3	**9**	47	0	0	0
3	0	0	2	**262**	2	80	1
4	0	0	0	0	**0**	0	0
5	0	0	0	8	0	**1**	0
6	3	0	0	0	72	10	**144**

	0	1	2	3	4	5	6
Precision	0.92	0.33	0.56	0.7	0.0	0.01	0.97
Recall	0.95	0.03	0.07	0.76	-	0.11	0.63
F1-Score	0.93	0.05	0.13	0.72	-	0.02	0.76

CW (ACC 67%)

True Class	0	1	2	3	4	5	6
0	**2678**	9	24	201	35	2	7
1	316	**7**	3	17	0	0	0
2	432	17	**18**	170	0	0	0
3	24	0	7	**609**	6	49	0
4	0	0	0	0	**0**	0	0
5	0	0	0	173	4	**36**	8
6	3	0	0	47	129	6	**46**

	0	1	2	3	4	5	6
Precision	0.78	0.21	0.35	0.5	0.0	0.39	0.75
Recall	0.91	0.02	0.03	0.88	-	0.16	0.2
F1-Score	0.84	0.04	0.05	0.64	-	0.23	0.32

grader 1 (ACC 71%)

True Class	0	1	2	3	4	5	6
0	**2158**	3	11	138	24	1	6
1	198	**3**	3	7	0	0	0
2	263	7	**10**	101	0	0	0
3	13	0	4	**472**	4	28	0
4	0	0	0	0	**0**	0	0
5	0	0	0	92	2	**17**	4
6	3	0	0	47	125	7	**48**

	0	1	2	3	4	5	6
Precision	0.82	0.23	0.36	0.55	0.0	0.32	0.83
Recall	0.92	0.01	0.03	0.91	-	0.15	0.21
F1-Score	0.87	0.03	0.05	0.69	-	0.2	0.33

grader 2 (ACC 57%)

True Class	0	1	2	3	4	5	6
0	**1764**	14	50	50	19	2	1
1	428	**18**	34	7	8	0	0
2	267	24	**53**	50	5	2	0
3	97	8	71	**133**	8	220	0
4	1	0	0	8	**39**	107	4
5	1	1	0	6	4	**13**	0
6	3	0	0	0	118	8	**153**

	0	1	2	3	4	5	6
Precision	0.69	0.28	0.25	0.52	0.19	0.04	0.97
Recall	0.93	0.04	0.13	0.25	0.25	0.52	0.54
F1-Score	0.79	0.06	0.17	0.34	0.22	0.07	0.7

Table 7.5: Confusion matrix for models using all features. The precision, recall, and F1-Score are calculated for each class.

7.6 Data visualization

Neonatologists using EEG to monitor the cerebral recovery of newborns are interested in their EEG transition over time in order to determine the health of the brain. While an automated grading system aids in interpreting the EEG patterns, a visualization of the predicted grades with a confidence output from the classifier helps physicians to gauge how much to trust the prediction, which ultimately serves to support the physician's decision making.

To visualize the classifier outputs in a form that allows for the intuitive interpretation of the predicted grades, as well as to see how confident we can be about the predictions, we created a heatmap with colors corresponding to the posterior probabilities from the SVM classifiers. Figures 7.1a–c and 7.2a–c show several visualization examples that display the results of patient 19 (chosen since it was the first case in the test set, not because of the classifier's predictive accuracy). The visualizations show the cerebral activity over approximately 29 hours of EEG recording, with both predictive and true class labels. This patient's EEG recording has only grades 0 – 3.

The visualizations in Figures 7.1a–d provide information about the confidence of the predictions by showing the probabilities of other grades. The red dots indicate **true** class

consensus (ACC 82%)

True Class	0	1	2	3	4	5	6
0	**1642**	0	6	57	19	0	2
1	78	**0**	1	0	0	0	0
2	82	0	**1**	42	0	0	0
3	2	0	1	**266**	3	75	0
4	0	0	0	0	**0**	0	0
5	0	0	0	6	0	**3**	0
6	0	0	0	0	73	11	**145**

	0	1	2	3	4	5	6
Precision	0.91	-	0.11	0.72	0.0	0.03	0.99
Recall	0.95	0.0	0.01	0.77	-	0.33	0.63
F1-Score	0.93	-	0.01	0.74	-	0.06	0.77

CW (ACC 67%)

True Class	0	1	2	3	4	5	6
0	**2653**	8	22	226	47	0	0
1	308	**7**	4	24	0	0	0
2	411	11	**22**	193	0	0	0
3	21	0	9	**625**	7	31	2
4	0	0	0	0	**0**	0	0
5	0	0	0	185	4	**22**	10
6	2	0	0	46	120	4	**59**

	0	1	2	3	4	5	6
Precision	0.78	0.27	0.39	0.48	0.0	0.39	0.83
Recall	0.9	0.02	0.03	0.9	-	0.1	0.26
F1-Score	0.84	0.04	0.06	0.63	-	0.16	0.39

grader 1 (ACC 71%)

True Class	0	1	2	3	4	5	6
0	**2121**	0	15	179	26	0	0
1	192	**4**	2	13	0	0	0
2	247	5	**14**	115	0	0	0
3	14	0	5	**484**	5	12	1
4	0	0	0	0	**0**	0	0
5	0	0	0	94	2	**16**	3
6	2	0	0	51	121	2	**54**

	0	1	2	3	4	5	6
Precision	0.82	0.44	0.39	0.52	0.0	0.53	0.93
Recall	0.91	0.02	0.04	0.93	-	0.14	0.23
F1-Score	0.86	0.04	0.07	0.66	-	0.22	0.38

grader 2 (ACC 59%)

True Class	0	1	2	3	4	5	6
0	**1720**	37	98	19	22	4	0
1	389	**60**	34	2	10	0	0
2	213	68	**87**	25	7	1	0
3	65	39	79	**159**	11	184	0
4	1	0	5	29	**36**	79	9
5	0	0	1	6	6	**12**	0
6	1	0	0	0	125	6	**150**

	0	1	2	3	4	5	6
Precision	0.72	0.29	0.29	0.66	0.17	0.04	0.94
Recall	0.91	0.12	0.22	0.3	0.23	0.48	0.53
F1-Score	0.8	0.17	0.25	0.41	0.19	0.08	0.68

Table 7.6: Confusion matrix for models using selected features. The precision, recall, and F1-Score are calculated for each class.

labels, and the colored blocks indicate the posterior probability of the predicted class, where pink is the highest probability and represents the predicted class. Figures 7.1a–b depict the *consensus* model, Figure 7.1c depicts the *grader 1* model, Figure 7.1d depicts the *grader 2* model. Note there are less hours displayed in Figures 7.1a–b than Figures 7.1c–d since only selected epochs in consensus were used.

We can observe that there are natural grade jumps due to SWC between grades within a short period of 5 minutes, so to display the overall EEG transitions with reduced noise, it is useful to smooth the predictions over some time. Figures 7.1a–c were smoothed by a moving average of four 5-minute epochs (20 minutes) and the probabilities were then normalized between 0 and 1; each block in the colorbar indicates a probability increment of 0.1.

To better display the transition of the EEG grades over time, we also use a line plot to visualize the predictions in a minimal fashion. Figures 7.2a–c show more clearly that towards hour 14 in *consensus*, the EEG became very non-stationary, and the classifier was able to capture this dynamic. Figure 7.1a corresponds to Figure 7.2a where both *consensus* models are displayed without any smoothing; similarly, Figure 7.1b depicts the *consensus* model with smoothing by 20 minutes, and corresponds to Figure 7.2b. Figure 7.2c is smoothed by 60 minutes, which further reduces noise.

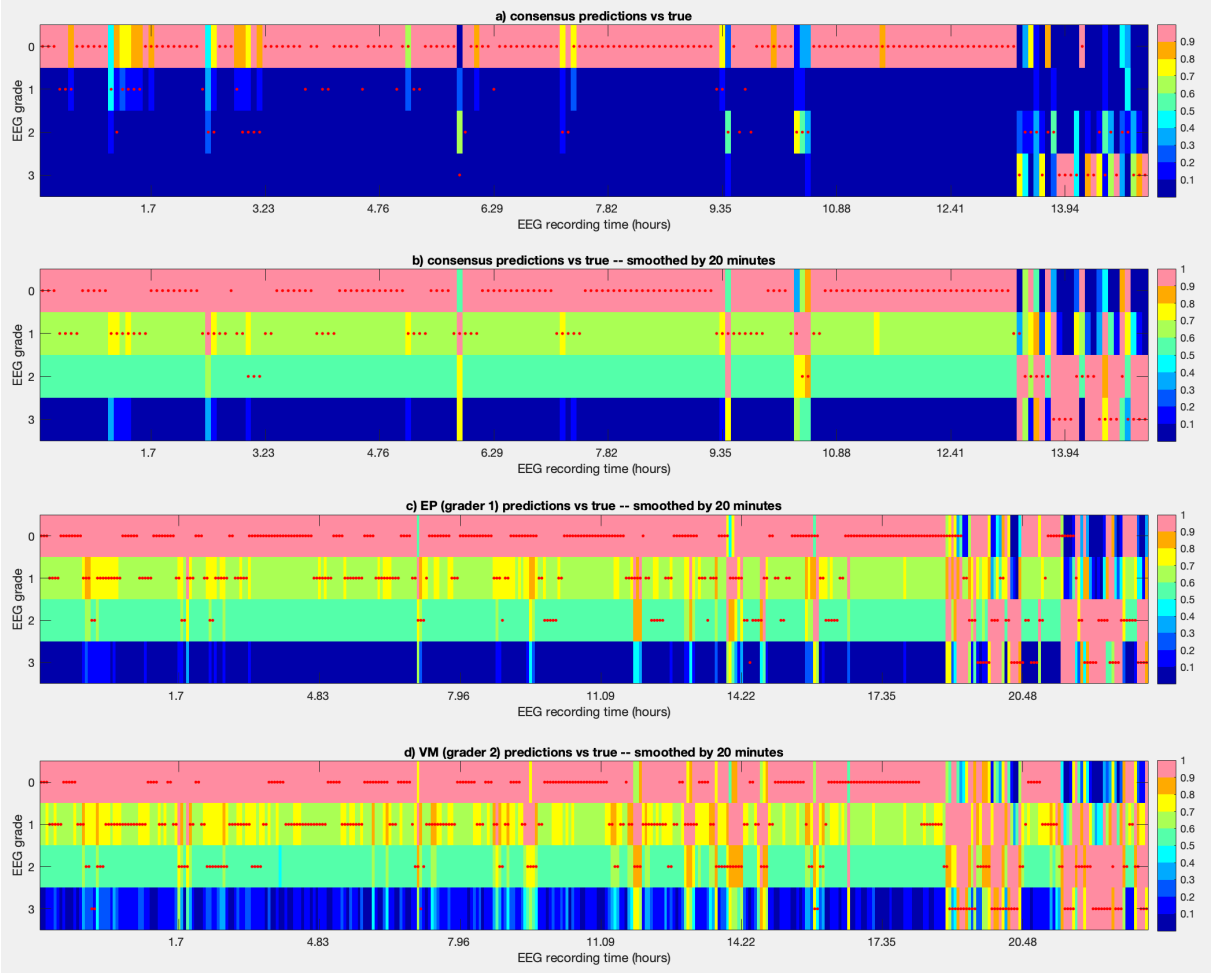

Figure 7.1: Visualization plots for patient 19, whose EEG recording displayed only grades 0 – 3. The red dots indicate **true** class labels, and the colored blocks indicate the posterior probability of the predicted class, where pink is the highest probability and represents the predicted class. Figures 7.1a–b depict the *consensus* model, Figure 7.1c depicts the *grader 1* model, Figure 7.1d depicts the *grader 2* model. Note that **a**–**b** display shorter hours than **c**–**d** since only selected epochs in consensus were used. **b**–**d** were smoothed by a moving average of four 5-minute epochs (20 minutes) and the probabilities were then normalized between 0 and 1; each block in the colorbar indicates a probability increment of 0.1. The visualization plots reflect the degree of variation between the graders' labels, and the classifiers were largely able to predict the graders' labels with high/moderate probability.

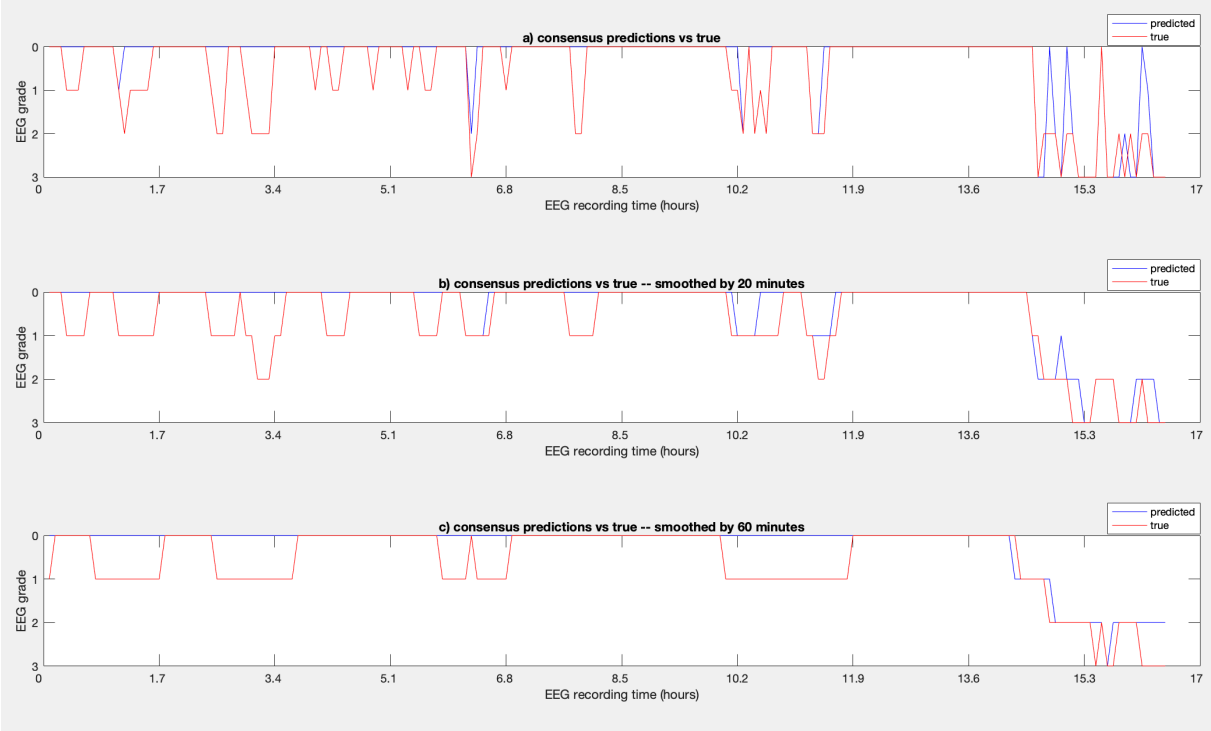

Figure 7.2: Visualization plots for patient 19, whose EEG recording displayed only grades 0 – 3. The blue lines indicate the predicted class labels, and the red lines indicate the true class labels. **a** shows the result of the *consensus* model without any smoothing, and **b** and **c** smooth the results by 20 and 60 minutes respectively. The overall EEG trend is displayed more clearly when noise is reduced with the smoothing.

8. Discussion

This thesis explores the relationship between inter-rater variability in the EEG grades and the classifier output. We are particularly interested in reconciling the manual grading of EEG signals by independent graders. The experiments in creating classifiers by using different graders' EEG grades as class labels were to ascertain how reliable a classifier can be in predicting the grades when there is ambiguity in the truth of the class labels. The ambiguity in the truth of the class labels mostly originated from the disagreement between the two graders on some EEG patterns.

The disagreements between the graders may have stemmed from human error and individual variability when visually interpreting the EEG signals. Upon evaluating an EEG trace, one grader may grade more conservatively than the other when uncertain about that instance. For example, Table 6.3 reveals that Grader 2 tends to classify some epochs as grade 5 when Grader 1 classified them as grade 3. The classifier using *grader 2*'s class label mirrored this disagreement and the grade 5 was often misclassified as 3. As the classifier performance was higher with *grader 1*'s class labels, we can speculate that perhaps the classifier had learned some features that are related to the features that Grader 1 visually interpreted during the grading.

In addition, the interpretation of EEG background activity may vary by nomenclature from different institutions of training, which may lead to variability between independent graders when interpreting the patterns. The two graders of the BABA dataset were trained at institutions in different countries, which may have contributed to the grading differences in the dataset.

The disagreements may also be related to signals that did not contain attributes that fell strictly under the definitions as described by the scoring system. Moreover, the EEG signals may also have contained some noise that were not filtered out during pre-processing, causing the signals to have less distinguishable features to discriminate against. For example, higher amplitude artifacts could mask lower amplitude activity and distort its pattern. If the EEG signals do not contain distinctive features between different grades as described in the scoring system to discriminate against, then both human and machine may confuse one class with another. Observing Figure 6.3 reveals that the graders mostly disagreed on grades that were close to each other or when the

EEG signals look visually similar.

Moreover, the scoring terminology may have contributed to the variations between the two graders' scores. As seen in Shellhaas et al's assessment, a simpler scoring system resulted in a higher inter-rater agreement than the more advanced scoring system did [28]. The Toronto scoring system is similar to the advanced scoring system with strict definitions for the EEG patterns, which may account for the moderate inter-rater agreement.

The epoch length used to grade the EEG patterns may also have contributed to the ambiguity of visual interpretation. EEG signals are non-stationary and can evolve over time, so when the graders evaluated the EEG at shorter epochs (e.g. 5 minutes), they may be have been assessing patterns in transition, which may present itself as sharing features similar to neighboring grades, or as if it could fit the definition of multiple EEG grades. Thus, there may be no clear answer to which grade of abnormality the pattern belongs to, contributing to grader disagreement.

We trained the *consensus* model with selected epochs that were graded in consensus and used it to evaluate the full-duration recording EEG data with the labels of Grader 1 and Grader 2. In comparison to their respective baseline models as reference, the *consensus* classifier did not outperform the baseline models, though the performance did not decrease. The *consensus* model was not able to generalize its learned features to classify the features of epochs graded in disagreement, which may be due to truly ambiguous features that are markedly different from other features due to noise.

In contrast, the multi-label classifier directly learned the ambiguous features with the full-duration recording, and its performance was comparable to the other models. Its overall accuracy was, as expected, lower than *consensus* since it was a more difficult task for the classifier to learn several labels for the same feature, rather than only one label. Moreover, it misclassified on average more patterns with severe abnormalities as mild/moderate than *consensus*, which is arguably a very undesirable property of the classifier.

To predict the outcome of a newborn with encephalopathy, it is reasonable to expect the classifier to minimize the number of false positives to the healthy scores so that the patients can receive the proper treatment. Similarly, it is undesirable for the classifier to have false positives to abnormal scores when it is truly healthy to prevent unnecessary treatment. In our experiments, all the classifiers rarely falsely classified severely abnormal patterns as healthy patterns. From the combined evaluation, the classifiers rarely predicted healthy patterns as moderate/severely abnormal patterns and vice versa.

The results of the classifiers reflect the properties of the EEG data. For all the models, grade 0 consistently obtained high precision and recall. We suspect that this was due to the class imbalance in the datasets, and was reflected in the performance of the

classifier where it frequently misclassified the grades 1 and 2 as 0, even if it is *not* a neighbor class. Since SVM's loss function attempts to maximize overall performance during training, it may have allowed misclassifications of the rarer classes to better discriminate between the larger classes when determining the margins.

The imbalanced dataset also affected the average accuracy in that it was different than the overall accuracy since the number of instances in each class is different. We can observe that the average accuracy was consistently higher than the overall accuracy since the classifiers could more easily detect grade 0 patterns, and there were more instances of grade 0 than other grades.

The large number of grade 0 (healthy, continuous pattern) in the data is reasonable since EEG signals are not static, and neonates who are recovering from HIE will naturally progress into more continuous EEG periods. Moreover, there are also natural variations in the patterns from the sleep-wake stages. Since cycling between states do not occur in higher grades of abnormality (major/inactive) than it does for normal/mild/moderate abnormalities, it is reasonable to see more ambiguity between grades 0 – 3.

Using feature selection resulted in the same overall classification accuracies, which suggests that only several features contribute to the classifier. Interestingly, most models selected $C = 8$ and $\gamma = 8$ as the best parameter with 5-fold cross-validation, other than *grader 2*. Recall that a smaller C allows for looser constraints and a larger margin, thus resulting in smaller penalty when assigning slack variables while optimizing the decision boundary allowing for more instances to cross the margins. This suggests that there was more difficult or ambiguous feature instances to distinguish between when classifying *grader 2*'s class labels.

The results of this thesis support the idea that automatic grading systems can be used as decision support systems to physicians in interpreting the EEG patterns, though they are not reliable enough as standalone interpreters. Altogether, the visual interpretation of EEG activity may lead to ambiguity in the EEG grades. While it does not seem to be a bottleneck in determining general findings (e.g. whether EEG abnormality grades have good prognostic value or not), it affects how effective peer review on EEG studies can be. Further studies of what classification schemes work best internationally will aid in developing a truly standardized system, and will also aid in developing the ideal automated grading system.

8.1 Future work

When assessing EEG background activity, it is useful to consider what time resolution is needed in the clinical case to evaluate the classifier. The physicians at BABA are interested in longer, overall EEG trends at 20 – 60-minute epochs, so we smoothed out the

5-minute epochs at a post-processing step for evaluation. Thus, there is an opportunity to incorporate the features of the shorter epochs during training to output hour-long epoch grade predictions. This would subsequently require a human grader to grade the EEG data by the hour.

We also did not have access to the raw EEG patterns in this thesis – but in the future we can incorporate pre-processing and calculating the EEG features to the automated grading system. In addition, future work include applying more advanced classification techniques, such as neural networks and other discriminative classifiers, and incorporating sleep-state cycling to adjust the predictions.

Since the multi-label classifier performed comparably to *consensus* and the reference models, it encourages efforts to optimize its performance. Some advantages of the multi-label classifier are that it can evaluate the full-duration recording instead of selected epochs, and it also reconciles inter-rater variability of the EEG graders by using both their scores as training data. We observed that the classifier cannot capture abnormal patterns as well as healthy patterns, which may be improved in the future by more severely penalizing false negatives to abnormal patterns during training.

Lastly, to further refine and evaluate the classifier, we could follow up with the newborns with confirmed HIE diagnosis and determine how well the manual grades and the classifier grades correlate with the outcome.

9. Conclusion

Using multi-class and multi-label classification with support vector machines, neonatal background EEG signals can be classified into seven scores of abnormality patterns. The work in this thesis also provides a method of visualizing EEG grades in full-term infants which can assist in monitoring the brain recovery of newborns who have suffered a HIE insult.

9.1 Acknowledgements

Thanks to Drs. Jukka Kohonen and Teemu Roos for supervising this Master's thesis; Dr. Sampsa Vanhatalo and the folks at BABA center for supervising and providing valuable medical insight for the project; Drs. Viviana Marchi and Elana Pinchefsky for labeling the EEG data; Ms. Minna Kauppila and Ms. Karoliina Tapani for pre-processing and calculating the data and features, and the rest of the folks involved in the NOGIN trial for the dataset.

References

[1] R. Ahmed, A. Temko, W. Marnane, G. Boylan, and G. Lightbody. Grading brain injury in neonatal EEG using SVM and supervector kernel. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5894–5898. IEEE, 2014.

[2] al Naqeeb N, A. D. Edwards, F. M. Cowan, D. Azzopardi, et al. Assessment of neonatal encephalopathy by amplitude-integrated electroencephalography. *Pediatrics*, 103(6 Pt 1):1263–1271, 1999.

[3] M. F. Bear, B. W. Connors, and M. A. Paradiso. *Neuroscience*, volume 2. Lippincott Williams & Wilkins, 2007.

[4] A. Beygelzimer, H. Daumé, J. Langford, and P. Mineiro. Learning reductions that really work. *Proceedings of the IEEE*, 104(1):136–147, 2016.

[5] J. Britton, L. Frey, and J. Hopp. *Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants.* American Epilepsy Society, 2016.

[6] T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1994.

[7] M. El-Dib, T. Chang, T. N. Tsuchida, and R. R. Clancy. Amplitude-integrated electroencephalography in neonates. *Pediatric Neurology*, 41(5):315–326, 2009.

[8] N. Ghamrawi and A. McCallum. Collective multi-label classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 195–200. ACM, 2005.

[9] E. M. Graham, K. A. Ruis, A. L. Hartman, F. J. Northington, and H. E. Fox. A systematic review of the role of intrapartum hypoxia-ischemia in the causation of neonatal encephalopathy. *American Journal of Obstetrics and Gynecology*, 199(6):587–595, 2008.

[10] L. Hellström-Westas, I. Rosén, L. De Vries, and G. Greisen. Amplitude-integrated aEEG classification and interpretation in preterm and term infants. *NeoReviews*, 7(2):e76–e87, 2006.

[11] C.-W. Hsu, C.-C. Chang, C.-J. Lin, et al. A practical guide to support vector classification. 2003.

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

[13] R. C. Knickmeyer, S. Gouttard, C. Kang, D. Evans, K. Wilber, J. K. Smith, R. M. Hamer, W. Lin, G. Gerig, and J. H. Gilmore. A structural MRI study of human brain development from birth to 2 years. *Journal of Neuroscience*, 28(47):12176–12182, 2008.

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

[15] J. J. Kurinczuk, M. White-Koning, and N. Badawi. Epidemiology of neonatal encephalopathy and hypoxic–ischaemic encephalopathy. *Early Human Development*, 86(6):329–338, 2010.

[16] M.-D. Lamblin, E. W. Esquivel, and M. André. The electroencephalogram of the full-term newborn: Review of normal features and hypoxic-ischemic encephalopathy patterns. *Neurophysiologie Clinique/Clinical Neurophysiology*, 43(5):267 – 287, 2013.

[17] J. Löfhede, N. Löfgren, M. Thordstein, A. Flisberg, I. Kjellmer, and K. Lindecrantz. Classification of burst and suppression in the neonatal electroencephalogram. *Journal of Neural Engineering*, 5(4):402, 2008.

[18] A. C. Lorena, A. C. De Carvalho, and J. M. Gama. A review on the combination of binary classifiers in multiclass problems. *Artificial Intelligence Review*, 30(1-4):19, 2008.

[19] R. Mani, H. Arif, L. J. Hirsch, E. E. Gerard, and S. M. LaRoche. Interrater reliability of ICU EEG research terminology. *Journal of Clinical Neurophysiology*, 29(3):203–212, 2012.

[20] V. Matić, P. J. Cherian, K. Jansen, N. Koolen, G. Naulaers, R. M. Swarte, P. Govaert, S. Van Huffel, and M. De Vos. Improving reliability of monitoring background EEG dynamics in asphyxiated infants. *IEEE Transactions on Biomedical Engineering*, 63(5):973–983, May 2016.

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

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

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

[24] P. L. Nunez and B. A. Cutillo. *Neocortical Dynamics and Human EEG Rhythms*. Oxford University Press, USA, 1995.

[25] J. Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999.

[26] R. Pressler, G. Boylan, M. Morton, C. Binnie, and J. Rennie. Early serial EEG in hypoxic ischaemic encephalopathy. *Clinical Neurophysiology*, 112(1):31–37, 2001.

[27] E. Shany, E. Goldstein, S. Khvatskin, M. D. Friger, N. Heiman, M. Goldstein, M. Karplus, and A. Galil. Predictive value of amplitude-integrated electroencephalography pattern and voltage in asphyxiated term infants. *Pediatric Neurology*, 35(5):335–342, 2006.

[28] R. A. Shellhaas, P. R. Gallagher, and R. R. Clancy. Assessment of neonatal electroencephalography (EEG) background by conventional and two amplitude-integrated aEEG classification systems. *The Journal of Pediatrics*, 153(3):369–374, 2008.

[29] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

[30] M. S. Sorower. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, 18:1–25, 2010.

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

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

[33] Y. Sun, A. K. Wong, and M. S. Kamel. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04):687–719, 2009.

[34] J. M. Toole and G. B. Boylan. NEURAL: Quantitative features for newborn EEG using Matlab. *arXiv preprint arXiv:1704.05694*, 2017.

[35] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[36] K. Tufenkjian. EEG instrumentation, montage, polarity, and localization. In *Epilepsy Board Review*, pages 15–32. Springer, 2017.

[37] J.-P. Vert, K. Tsuda, and B. Schölkopf. A primer on kernel methods. *Kernel Methods in Computational Biology*, 47:35–70, 2004.

[38] A. J. Viera, J. M. Garrett, et al. Understanding interobserver agreement: The kappa statistic. *Fam Med*, 37(5):360–363, 2005.

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

[40] K. Watanabe, S. Miyazaki, K. Hara, and S. Hakamada. Behavioral state cycles, background EEGs and prognosis of newborns with perinatal hypoxia. *Electroencephalography and Clinical Neurophysiology*, 49(5-6):618–625, 1980.

[41] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

Appendix A. Calculated EEG Features

EEG features		
amp_mean	mobility	band10_12
amp_variance	complexity	band12_30
amp_skew	Nonlinear_energy	band0_2norm
amp_kurt	H_spectral	band1_3norm
freq_mean	Zero_crossings	band2_4norm
freq_variance	AR1	band3_5norm
freq_skew	AR2	band4_6norm
freq_kurt	AR3	band5_7norm
Cov_IA_IF	AR4	band6_8norm
fractal_dim	AR5	band7_9norm
dfa	AR6	band8_10norm
mdfa	AR7	band9_11norm
mdfa_max	AR8	band10_12norm
rEEG_5	AR9	band12_30norm
m_med	total_power	SEF90
burst_duration	band0_2	SEF95
burst_nro	band1_3	SEF80
ibi	band2_4	kurt
Line_length	band3_5	skew
wavelet_energy	band4_6	svd_entropy
min_max	band5_7	fisher
RMS_amp	band6_8	ZC1d
H_shannon	band7_9	ZC2d
peak_freq	band8_10	V1d
activity	band9_11	V2d