Ollscoil na hÉireann, Corcaigh

National University of Ireland, Cork

Neonatal seizure detection: A deep learning approach

Thesis presented by Alison O'Shea

for the degree of Doctor of Philosophy

University College Cork

School of Engineering

Head of School/Department: Prof. Jorge Oliveira

Supervisors: Dr Gordon Lightbody & Dr Andriy Temko

September 2020

Table of contents

Declaration

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism and intellectual property.

Alison O'Shea

Acknowledgements

Meitheal represents a community coming together to achieve something positive; this thesis and the years of work that it represents was undoubtedly a collective effort.

I want to acknowledge the support of my supervisors Dr Andriy Temko and Dr Gordon Lightbody. Many years ago, they allowed me to take on this project, and since then they have supported me through the fun times and the tougher days. It was a pleasure to learn from them and I hope to emulate their brilliance and patience in all my future endeavours.

I want to thank the larger ecosystems which facilitated my scholarship. Firstly, the INFANT research centre and all the wonderful people which contribute to its success. Being part of a diverse centre was a vital part of my PhD experience and I want to thank all the friends, colleagues, and mentors that I met through INFANT, especially Prof. Geraldine Boylan for her support. Secondly, I am eternally grateful to everyone who works in the electrical engineering building in UCC, in all capacities. This has been my second home for almost 10 years, I would not be at this stage without these outstanding people who I have had the pleasure of working with along the way. A special mention goes to Ralph and Niamh who were always willing to lend a listening ear and to help with any request no matter how big or small.

I was extremely lucky to be surrounded by an amazing cohort of peers over the last five years. There was a constant flow of students through the building, but one thing that was common among everyone was the positivity and unquestionable support that they offered. I need to specifically mention Brian, Oksana, and Mark; together we navigated the worlds of engineering and medicine, and I would not have made it without your continued friendship.

I am deeply grateful to my family and friends and for their unwavering encouragement and belief in me. My parents Michael and Irene, and my brother Daire epitomise kindness, intelligence, and selflessness; each day I learn from them and cherish their support in all aspects of my life. Finally, I want to extend my sincere thanks to James, who was with me through every step of this journey; thank you for listening when I needed to rant about work and accepting when I refused want to discuss it. Any remaining typos will be blamed on you!

Thank you all for being part of my team, this work is as much yours as it is mine.

Buíochas ó chroí libh

Abstract

The detection of neonatal seizures is an important step in identifying neurological dysfunction in newborn infants. Research indicates that seizures in infants are underreported; specialist EEG monitoring equipment and neonatal neurophysiological expertise are required to reliably detect neonatal seizures due to the predominance of sub-clinical seizure events. Often, the required level of clinical expertise is not available around-the-clock in the NICU; medical experts have cited the need for decision support algorithms to assist staff during these times.

In this thesis novel automated neonatal seizure detection algorithms are proposed. Here deep learning end-to-end optimised algorithms, which detect seizures from raw multi-channel EEG, are presented. The deep learning approach differs from previous rule-based and machine learning systems, which rely on hand-engineered features. An appropriate input EEG representation is selected through empirical analysis; 2-dimensional time-frequency representations of EEG are compared to 1-dimensional temporal representations in a series of experiments. Deep learning algorithms prove capable of extracting discriminative intermediate hierarchical representations from raw EEG.

A deep learning algorithm for detecting seizures in term neonates is proposed. The designed algorithm utilises only convolutional layers to process multi-channel temporal EEG and is designed to exploit the large quantity of weakly labelled data available in the training stage. The effect of varying architectural parameters is thoroughly studied, and the designed architecture compares favourably in terms of performance, inference run time and number of parameters when compared with baseline systems. The developed system outperforms state-of-the-art machine learning algorithms when tested on a large database of continuous EEG recordings (duration 834h) and when further validated on a held-out publicly available dataset (duration 112h), achieving AUC results of 98.5% and 95.6% respectively.

The challenges associated with detecting seizures in term EEG are exacerbated in preterm EEG due to the large variations in EEG morphology depending on gestational age and the limited amount of annotated preterm EEG available. A deep learning algorithm development framework is proposed where classifiers are trained to detect seizures for specific gestational age ranges. The developed algorithms leverage the existence of robustly trained term EEG models through transfer learning and classifier ensembling; this results in accurate seizure detection despite the scarcity of labelled training data. This preterm seizure detection

framework represents the first time an algorithm of this kind has been developed; it achieves an AUC of 95.4% on a held-out test dataset (duration 575h), and detects approximately 50% of all seizure events at a false detection rate of one false alarm every four hours.

In this work, algorithms are investigated through a series of visualisation techniques. This analysis gives an understanding of the EEG patterns which contribute to algorithm decisions and highlights the differences between term and preterm EEG. The developed algorithm framework is also utilised as part of a mobile brain monitoring device where the light-weight nature of the designed network and the simplicity of the inference computations are exploited through AI-on-the-edge decision support. The ability to decipher deep learning algorithms and to integrate algorithms into existing brain monitoring systems are necessary steps for the translation of the work in this thesis into clinical domain.

List of figures

List of tables

List of acronyms

aEEG	**A**mplitude integrated **EEG**
AI	**A**rtificial **i**ntelligence
ANSeR	**A**lgorithm for **n**eonatal **se**izure **r**ecognition
ASIC	**A**pplication **s**pecific **i**ntegrated **c**ircuit
AUC	**A**rea **u**nder the ROC **c**urve
BCI	**B**rain **c**omputer **i**nterface
CNN	**C**onvolutional **n**eural **n**etwork
CPU	**C**omputational **p**rocessing **u**nit
CUMH	**C**ork **U**niversity **M**aternity **H**ospital
cUS	**C**ranial **u**ltra**s**ound
ECG	**E**lectro**c**ardio**g**ram
EEG	**E**lectro**e**ncephalo**g**ram
FCN	**F**ully **c**onvolutional **n**etwork
FD/h	**F**alse **d**etections per **h**our
FFT	**F**ast **F**ourier **t**ransform
FLOP	**F**loating **p**oint **op**eration
GA	**G**estational **a**ge
GDR	**G**ood **d**etection **r**ate
GMM	**G**aussian **m**ixture **m**odel
GPU	**G**raphical **p**rocessing **u**nit
HIE	**H**ypoxic-**i**schemic **e**ncephalopathy
IOA	**I**nter-**o**bserver **a**greement
IVH	**I**nter**v**entricular **h**aemorrhage
KNN	**k**-**n**earest **n**eighbours
LARS	**L**ayer-wise **a**daptive **l**earning **r**ate **s**caling
LOO	**L**eave **o**ne **o**ut

LSTM	**L**ong **s**hort-**t**erm **m**emory
MLP	**M**ulti-**l**ayer **p**erceptron
MRI	**M**agnetic **r**esonance **i**maging
NICU	**N**eonatal **i**ntensive **c**are **u**nit
NIRS	**N**ear-**i**nfra**r**ed **s**pectroscopy
NLEO	**N**on-**l**inear **e**nergy **o**perator
NN	**N**eural **n**etwork
P-SDA	**P**reterm **s**eizure **d**etection **a**lgorithm
ReLU	**Re**ctified **l**inear **u**nit
RF	**R**eceptive **f**ield
RNN	**R**ecurrent **n**eural **n**etwork
ROC	**R**eceiver **o**perating **c**haracteristic
SC	**S**pike **c**orrelation
SGD	**S**tochastic **g**radient **d**escent
SVM	**S**upport **v**ector **m**achine
TFC	**T**ime **f**requency **c**orrelation
TL	**T**ransfer **l**earning
TPU	**T**ensor **p**rocessing **u**nit
T-SDA	**T**erm **s**eizure **d**etection **a**lgorithm

Chapter 1: Introduction

Detecting neonatal seizures is an important and complex healthcare challenge. During the gestational and neonatal stages, the brain undergoes intense development and structural changes; any problems associated with this vital organ can have critical effects on survival and quality of life (Uria-Avellanal et al., 2013). Although mortality following neonatal seizures is decreasing, the rate of long-term adverse effects has remained at approximately 30% following seizures in full term infants and it remains even higher in preterm infants (Silverstein & Jensen, 2007). Long term adverse effects include postneonatal epilepsy, cerebral palsy, and reduced intelligence quotient at 7 years, with higher incidences reported following seizures in preterm populations (Ramantani et al., 2019; Yıldız et al., 2012). Neonatal seizures are often the first indicator of underlying brain injury. Timely detection of seizures allows clinicians to diagnose significant problems early and gives the neonatal intensive care unit (NICU) staff the opportunity to administer anti-epileptic drugs and treatments at the appropriate time. Current rates of seizure detection, in even the most well-resourced NICUs, remain low. One study showed that as little as 9% of seizure events in a NICU were recognised by clinical staff (Murray et al., 2008, Clancy et al., 1988).

The problem is made more difficult by the subtle characteristics of seizures in newborn infants; studies report around 66% of neonatal seizure burden has no clinical correlate; sub-clinical seizures occur when there are no overt clinical signs that an infant is experiencing a seizure event (Murray et al., 2008). In these instances, seizures are detectable through temporal monitoring of the brain. But there is considerable complexity associated with reading the electroencephalogram (EEG) waveform, which is widely accepted to be the gold-standard method of detecting seizures (Shellhaas, 2015). Neonatal neurophysiologists are highly trained clinical staff who specialise in interpreting neonatal EEG. These experts develop an understanding, and often an intuitive feel, for the correlation between specific patterns and different clinical diagnoses (Malone et al., 2009). Their work is vital in the NICU, they provide a deep level of understanding that currently no other diagnostic test or staff member can contribute. They study the characteristics present in the neonatal brain at different stages of development.

Electroencephalography refers to the recording and monitoring of electrical signals in the brain. The first recording of EEG in humans is attributed to Hans Berger in 1924 (Haas, 2003). In the

almost one hundred years since the first human recordings of EEG, the study of human electroencephalographic signals has become a vital medical tool. The analysis of EEG signals allows for the diagnosis of many brain disorders such as epilepsy and stroke and gives a deeper understanding of brain activity, which is utilised in diagnosing sleep disorders and monitoring patients under anaesthesia (Hori et al., 2001; Jameson & Sloan, 2006). The understanding and analysis of EEG signals is also a key component of brain-computer interface research (BCI) which focuses on the development of devices which can receive input directly from a user's brain without an intermediate input device, such as a mouse, keyboard or microphone (Wolpaw et al., 2000). The brain is the origin of all conscious and subconscious human actions, and EEG is an important tool for understanding this vital organ.

Some of the earliest research into artificial intelligence (AI) was influenced by the relative simplicity and computational power of biological neurons in the brain. Human brains appear to learn complex patterns by changing the strength of synaptic connections between networks of neurons. The 1940's-1960's saw the first theoretical research into artificial neurons and artificial neural networks (NN) and the development of early implementations of these theories (McCulloch & Pitts, 1943; Rosenblatt, 1958; Widrow & Hoff, 1960). The development of the perceptron and the associated learning algorithm was seen as an extremely promising step forward in the field of AI and especially in pattern recognition applications. This convex training algorithm is guaranteed to find a weight vector which satisfies all training cases, if such a solution exists. But, in practice, single layer perceptrons trained using the perceptron learning algorithm developed by McCulloch and Pitts had many limitations (Minsky & Papert, 1969). The classic example which shows the perceptron's limitations is that it cannot learn the relatively simple XOR function; the input vectors cannot be linearly separated in the input dataspace. The perceived limitations of NNs led to a decline in research in this area and for many years progress was slow. The emergence of the backpropagation algorithm, which could be applied to multi-layered non-linear networks of feed-forward neurons, prompted a resurgence in research into artificial NNs (Rumelhart et al., 1986). The inclusion of multiple layers and non-linearities meant that the newly developed multi-layer perceptrons did not have a convex loss function, therefore they were liable to get stuck in local minima during training.

During the time when NNs were considered difficult to train, the support vector machine (SVM) classifier was growing in popularity. Unlike NNs, SVMs have a convex loss function, so the algorithm is guaranteed to find the optimal solution. Artificial NNs were also shown to be suffering from 'vanishing gradients'; even with the newly developed backpropagation

algorithm the gradient updates become increasingly small in the layers furthest from the output due to activation function saturation. We now know that recent advances in hardware, algorithms and the increased availability of training data have given rise to success in training deep NNs, but in the early 20th century the SVM was a more mathematically robust, less data hungry, and more computationally efficient alternative.

The soft margin classifier implementation of the SVM algorithm was first published in 1995 (Cortes & Vapnik, 1995). The SVM is a maximum margin classifier, which is optimised to reduce the generalisation error of the classifier. The SVM is typically trained using quadratic programming and represents a convex constrained optimisation algorithm which is guaranteed to find the global optimum. The SVM employs a kernel trick which also means that non-linear feature spaces can be learned. Kernels offer an efficient way of transforming the data into a higher dimensional space, they aim to represent the data in a way that is linearly separable. SVM classifiers rely on the extraction of an informative feature set. SVMs, along with many other machine learning algorithms, learn from numeric feature vectors. Each feature vector represents the important characteristics of an input sample.

The NN renaissance came in 2012 when a deep learning algorithm won an international image processing challenge, ImageNet, by a significant margin (Krizhevsky et al., 2012). The magnitude of the performance improvement was unprecedented. It is interesting to note that the two best performing classifiers on ImageNet in the previous year, 2011, both utilised SVM classifiers (Sanchez & Perronnin, 2011; van de Sande et al., 2011). From 2012 onwards all of the winning entries in the ImageNet competition have been deep learning algorithms (Russakovsky et al., 2015). Deep learning and neural networks had been increasing in popularity in the decade preceding 2012's ImageNet competition, but this public benchmark of their success and the reutilization of convolutional network architectures (ConvNet or CNN) forced the general AI research community to take notice. CNNs could learn from untransformed input data and were not reliant on the extraction of an informed feature set.

The so-called AlexNet model which won the ImageNet competition in 2012 was named after first author Alex Krizhevsky, the algorithm was a reimplementation of an earlier model developed by Yann LeCun (Krizhevsky et al., 2012). LeCun's 1989 neural network used layers of shared weights followed by fully connected layers and all of the network weights were learned during training (LeCun et al., 1989). It was the confluence of three new ideas that made the AlexNet model so powerful. The first idea was that sigmoid non-linearities were replaced

with rectified linear units (ReLu), this reduced the effect of diminishing gradients in deep architectures (Nair & Hinton, 2010). The second inclusion was dropout in the fully connected layers; dropout forces a network to learn more robust internal features, as a different subset of nodes will be presented at each training iteration (Hinton et al., 2012). The third innovation was a hardware adaptation, where the neural network was trained on two graphical processing units (GPU), which yielded a significant increase in training speeds over training the network on a computational processing unit (CPU). GPUs are well suited to parallelised tasks. The combination of these three ideas resulted in the dawn of our current era of deep learning driven AI.

As early as the 1970's, when AI was still in its nascency, the healthcare sector was identified as one of the most promising application areas for AI systems (Yu et al., 2018). Since the mid-twentieth century, automated decision support systems have been a growing research area and commercial market year upon year. Their ability to assist in clinical diagnosis has influenced healthcare across the spectrum. There is currently no field where the final decision is removed from the human expert's hand, but it is predicted that the first case of machine selected treatment may come in the domains where diagnosis relies on detecting patterns in images. In the domains of radiology, dermoscopic melanoma diagnosis and fundus photograph evaluation, AI systems have reached expert level performances (Gulshan et al., 2016; Haenssle et al., 2018; Liew, 2018; Poplin et al., 2018; Rajpurkar et al., 2017). One thing these systems have in common is that they all rely on deep learning. Considering that deep learning systems require large datasets to train on, these advances would not have been possible without the vast quantities of medical information currently available in electronic health records. The level of physiological monitoring in medicine is what allows for the utilisation of deep learning in this big data field ('Ascent of Machine Learning in Medicine', 2019).

In the NICU, the development of devices to monitor signals and understand each infant's physiology has revolutionised neonatal healthcare (Rhine, 2016). Clinical staff utilise monitored data to create treatment plans and to understand when an infant requires immediate intervention. But experienced experts are not always available to interpret the various data sources. The variety and velocity of data being recorded in NICUs is an untapped resource which could be further exploited through the development of AI clinical decision support tools.

Intelligent EEG monitoring in the NICU, and, more specifically, automated neonatal seizure detection, has been an area of research for almost 30 years. The earliest algorithms sought to

characterise specific seizure signatures and to create features which could detect particular patterns (Gotman et al., 1997; Liu et al., 1992). The algorithms developed by Liu *et al.* (1992) and Gotman *et al.* (1997) both defined features, based on clinical knowledge of how neonatal seizures manifest in EEG and developed heuristic thresholds to differentiate between seizures and background activity. Another approach developed to recognise neonatal seizures measured the complexity of an EEG window (Celka & Colditz, 2002). This is based on the understanding that a neonatal seizure represents a more ordered and rhythmic event, when compared to background EEG windows. These early algorithms did not employ data-driven classifiers, instead thresholds and rules were hand-tuned by engineers. This leads to patient or dataset specific performances which lack the ability to generalise to unseen patients; an independent study of these three algorithms found that none achieved performances suitable for adoption within a NICU (Faul et al., 2005).

Over the next decade, the rise in popularity of machine learning led to the development of data driven models. These algorithms replaced the previous feature driven heuristic algorithms and instead employed large more generalised feature sets and replaced human engineering thresholds with machine learning algorithms (Ahmed et al., 2017; Tapani et al., 2019; Temko et al., 2011). The chosen feature sets still contained features which were driven by medical knowledge about the nature of seizure events, but these highly engineered features were supplemented with more general features driven by the domain of signal processing. Researchers are still searching for one sole feature that can be used to perfectly discriminate between seizure and non-seizure EEG windows. Since there is no one catch-all feature, machine learning can derive knowledge from the combinations of features that do work, or from the patterns in the EEG which correlate most with seizure detections.

One major success in this generation of machine learning based neonatal seizure detection algorithms was the development of the ANSeR algorithm (Temko et al., 2013, 2015). The ANSeR machine learning algorithm achieved an acceptable performance during research and development stages, and it has been further validated in a clinical trial with the eventual goal of becoming a clinically available decision support tool (Mathieson et al., 2016). The development of a commercial tool and its progression through clinical trials is an important milestone for AI in NICU decision support. The existence of a recognised commercial tool provided a benchmark for comparison against when developing new neonatal seizure detection algorithms.

Most of the previous work on developing automated seizure detection algorithms has fallen into one of two categories: engineering new features for the seizure detection task or using machine learning tricks to create more robust neonatal seizure classifiers. The motivation of this thesis is to look at the shortcomings of these two approaches and to develop a novel way of approaching the development of an automated seizure detection classifier. Deep learning algorithms do not rely on the extraction of an engineered feature set, instead task-specific features are learned as a part of an end-to-end classifier optimisation routine.

Deep learning has existed for a few decades but in recent years there has been a surge in interest in deep learning applications. The recent advance in deep learning is the culmination of three important developments: hardware, algorithms, and data. Through harnessing GPUs and parallel computing resources in general, deep learning engineers can speed up computations and can process more data for longer training times than ever before. The development of new algorithms for architecture regularisation and learning stabilisation, makes training more efficient. The increase in data generation and data storage have been exploited by deep learning algorithms, which require large training datasets from which to learn. These factors coupled together make deep learning a powerful new area of research that can be applied to the task of neonatal seizure detection.

1.1 The aims and scope of this thesis

The work in this thesis was conducted with the intention of developing and testing a deep learning based neonatal seizure detection algorithm. The goal was to develop an objective decision support tool to aid in seizure diagnosis. The algorithms developed in the field of automated seizure detection engineering are always intended as decision support tools, not diagnostic tools. The algorithm development utilised EEG signals collected in real NICU environments, no manual artefact rejection or removal was included, and no hand-crafted features were developed. No other environmental (video) or physiological (ECG, respiration) signals were employed to assist in seizure diagnosis. The intention was to improve the way EEG is utilised in the NICU.

A first aim is to understand if deep learning algorithms are capable of learning informative features which can be utilised to distinguish seizure events in EEG.

The second aim is to optimise a deep learning algorithm to improve on the current state-of-the-art neonatal seizure detection algorithms. The developed algorithms are compared with established machine learning based neonatal seizure detection algorithms in terms of algorithm

sensitivity, specificity, and the computational resources required to implement them. This work also seeks to understand the architectural parameters of the developed deep learning algorithms which impact on the performance. This involves creating a comparative framework where the characteristics of the deep learning architecture can be explored.

The third aim of this work is to establish whether the developed term seizure detection algorithm could be applied to preterm EEG to detect seizures. Preterm EEG presents a more complex problem than that of detecting seizures in term EEG. Preterm EEG varies with age and there are less annotated data available for algorithm training.

The fourth aim of this work is to show how the trained deep learning networks can be interrogated. Properties of the developed deep learning architecture are exploited to highlight input patterns which contribute to classifier decisions. The learned characteristics of trained networks hidden layers are analysed and age specific learned filters are contrasted.

The fifth aim of this work is to demonstrate the utility of deep learning on mobile platforms for real time clinical decision support. The computational resources and the performance trade-offs required to implement the algorithms on low-resource devices are outlined.

1.2 The main contributions of this thesis

This thesis represents the first work where raw EEG signals, with no intermediate feature representation, are used to train a classifier to detect neonatal seizures. The very foundation of this work is novel, the shortcomings of previous approaches were outlined, and deep learning was harnessed to over-come problems with data utilisation and the need for an engineered feature representation.

In particular, the main contributions of this thesis are:

1. A neural network was created and trained with raw EEG as input to avoid human biases, creating a purely data driven feature extractor
2. The novel architecture was further designed so that it does not require channel specific seizure annotations during training, this vastly increases the amount of training data available
3. The architecture trained on full term neonatal EEG was utilised through transfer learning to train a seizure detector for preterm neonatal EEG, despite having a very limited dataset of preterm EEG available

4. The features learned in the preterm and term models are visualised and contrasted to give an understanding of the differences in the data driven feature extraction stages
5. Trained deep learning algorithms for neonatal seizure detection are implemented on mobile devices and the performance is analysed from a computational and classification perspective

1.3 Thesis layout

The development of an automated neonatal seizure detection algorithm to provide decision support in the NICU requires multi-disciplinary research and input from a wide variety of specialist fields. The work encapsulated in this thesis draws on knowledge from the fields of neuroscience, medicine, engineering, and computer science. The layout of this thesis reflects the broad focus of this work and is structured as follows:

Chapter 2 details important background information about the characteristics of the neonatal brain and different methods of monitoring neonatal brain status. This chapter also provides information about neonatal seizures, how they manifest in the neonatal period, and methods for detecting seizure events. Some readers may not have experience in the area of neonatal seizure detection, therefore for those readers it is especially important to understand the clinical motivation and challenges associated with detecting neonatal seizures. This chapter also gives details about the different types of seizure labels utilised in this work. This work focuses on the development of a supervised neonatal seizure detection algorithm, so the availability of labels is an important aspect of algorithm training. Finally, Chapter 2 gives information about the different datasets of multi-channel EEG with annotated seizures utilised for training and testing in this work.

Chapter 3 presents a technical description of previously developed neonatal seizure detection algorithms. These include feature-based approaches which mimic the clinical diagnostic procedure and data-driven machine learning based approaches which rely on large feature sets combined with machine learning algorithms. Chapter 3 also details methods of quantifying algorithm performance and presents two algorithms which achieve reliable neonatal seizure detection performances, with AUCs above 95%, and will be utilised as benchmarks throughout this thesis. Finally, Chapter 3 presents an overview of deep learning and convolutional neural networks, which are the algorithms employed in this work.

Chapter 4 presents an explorative study of the suitability of deep learning algorithms for detecting seizures in unprocessed neonatal EEG. In this chapter, potential EEG representations

are defined, and preliminary experiments are conducted to compare the performance of simple deep learning architectures for various EEG input representations. This chapter also outlines the different classification architectures which can be utilised as part of a convolutional network. Finally, this chapter includes details about the framework utilised in this thesis to implement the deep learning algorithms from a hardware and a software perspective.

Chapter 5 follows on from the learnings in Chapter 4; deep learning algorithms can differentiate between seizure and non-seizure epochs of raw temporal EEG. This chapter builds on that by exploring the architectural parameters of a fully convolutional network which most influence the performance. An architecture which utilises weakly labelled seizure data for supervised learning of a deep convolutional network is also explored and the performance of the developed algorithm is compared to both the ANSeR and the Helsinki SVM algorithms.

Chapter 6 investigates the prospect of applying a deep learning algorithm, which was trained and validated on term EEG, to detect seizures in preterm EEG. Term and preterm EEG signals vary considerably, and features which are considered normal for one gestational age would be considered highly abnormal for others. This chapter also explores the development of neonatal seizure detection algorithms specifically for preterm seizure detection and explores how algorithm effectiveness depends on gestational age. The study tests and compares several approaches to address the problems associated with training on full term data, training on preterm data, training on age-specific preterm data, and transfer learning. The proposed deep learning approach avoids time-consuming explicit feature engineering and leverages the existence of the term seizure detection model, resulting in accurate predictions with a minimum amount of annotated preterm data.

Chapter 7 focuses on the translation of the networks developed in this work into a clinical setting. The interpretation of these data driven deep learning algorithms is explored through the analysis of inputs, visualisation of learned filers, and the synthesis of seizure and non-seizure EEG. Implementation of the network architecture on a low-memory, low-power platform which could provide clinical decision support in NICU environments is investigated. The performance of the networks developed during this thesis when realised on an off the shelf android tablet and a custom embedded system design are detailed.

Chapter 8 presents a concluding summary which reviews the findings of each chapter in this work. The advantages and disadvantages of the proposed deep learning framework are explored, and potential future research arising from this thesis is outlined.

1.4 Publications arising from this work

Journal publications

O'Shea, A., Lightbody, G., Boylan, G., & Temko, A., 2020. "Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture", *Neural Networks*, 123, pp. 12-25.

O'Shea, A., Ahmed, R., Lightbody, G., Mathieson, S., Pavlidis, E., Pisani, F., Marnane, W., Boylan, G., Temko, A., "Automated EEG-based seizure detection in preterm infants", *International Journal of Neural Systems*, In press.

O'Sullivan, M., **O'Shea, A.**, Power, A., Yang, B., Temko, A., Popovici, E., "Design of an edge AI device for neonatal EEG acquisition and interpretation", prepared.

Conference publications

O'Shea, A., Lightbody, G., Boylan, G., & Temko, A., 2018. Investigating the impact of CNN depth on neonatal seizure detection performance. Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5862-5865.

O'Sullivan, M., Gomez, S., **O'Shea, A.**, Salgado, E., Huillca, K., Mathieson, S., Boylan, G., Popovici, E. and Temko, A., 2018. Neonatal EEG interpretation and decision support framework for mobile platforms. Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4881-4884.

O'Shea, A., Lightbody, G., Boylan, G., Temko, A., 2017. Neonatal seizure detection using convolutional neural networks. IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1-6.

1.5 References

Ahmed, R., Temko, A., Marnane, W. P., Boylan, G., & Lightbody, G. (2017). Exploring temporal information in neonatal seizures using a dynamic time warping based SVM kernel. *Computers in Biology and Medicine*, *82*, 100–110.

Ascent of machine learning in medicine. (2019). *Nature Materials*, *18*(5), 407.

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

Clancy, R. R., Legido, A., & Lewis, D. (1988). Occult neonatal seizures. *Epilepsia*, *29(3)*, 256–261.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

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

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

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, *316*(22), 2402–2410.

Haas, L. (2003). Hans Berger (1873–1941), Richard Caton (1842–1926), and electroencephalography. *Journal of Neurology, Neurosurgery, and Psychiatry*, *74*(1), 9.

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A. B. H., Thomas, L., Enk, A., & Uhlmann, L. (2018). Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, *29*(8), 1836–1842.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *ArXiv:1207.0580 [Cs]*.

Hori, T., Sugita, Y., Koga, E., Shirakawa, S., Inoue, K., Uchida, S., Kuwahara, H., Kousaka, M., Kobayashi, T., Tsuji, Y., Terashima, M., Fukuda, K., Fukuda, N., & Sleep Computing Committee of the Japanese Society of Sleep Research Society. (2001). Proposed supplements and amendments to 'A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects', the Rechtschaffen & Kales (1968) standard. *Psychiatry and Clinical Neurosciences*, *55*(3), 305–310.

Jameson, L. C., & Sloan, T. B. (2006). Using EEG to monitor anesthesia drug effects during surgery. *Journal of Clinical Monitoring and Computing*, *20*(6), 445–472.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (pp. 1097–1105).

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, *1*(4), 541–551.

Liew, C. (2018). The future of radiology augmented with Artificial Intelligence: A strategy for success. *European Journal of Radiology*, *102*, 152–156.

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

Malone, A., Ryan, C. A., Fitzgerald, A., Burgoyne, L., Connolly, S., & Boylan, G. B. (2009). Interobserver agreement in neonatal seizure identification. *Epilepsia*, *50*(9), 2097–2101.

Mathieson, S., Rennie, J., Livingstone, V., Temko, A., Low, E., Pressler, R. M., & Boylan, G. B. (2016). In-depth performance analysis of an EEG based neonatal seizure detection algorithm. *Clinical Neurophysiology*, *127*(5), 2246–2256.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133.

Minsky, M., & Papert, S. (1969). Perceptrons. *MIT Press, Cambridge, MA*.

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

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML)*, 807–814.

Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., Peng, L., & Webster, D. R. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, *2*(3), 158–164.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *ArXiv:1711.05225 [Cs, Stat]*.

Ramantani, G., Schmitt, B., Plecko, B., Pressler, R. M., Wohlrab, G., Klebermass-Schrehof, K., Hagmann, C., Pisani, F., & Boylan, G. B. (2019). Neonatal Seizures—Are We there Yet? *Neuropediatrics*, *50*(5), 280–293.

Rhine, W. D. (2016). Technology – Considerations for the NICU of the Future. *Newborn and Infant Nursing Reviews*, *16*(4), 208–212.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Sanchez, J., & Perronnin, F. (2011). High-dimensional signature compression for large-scale image classification. *CVPR 2011*, 1665–1672.

Shellhaas, R. A. (2015). Continuous long-term electroencephalography: The gold standard for neonatal seizure diagnosis. *Seminars in Fetal and Neonatal Medicine*, *20*(3), 149–153.

Silverstein, F. S., & Jensen, F. E. (2007). Neonatal seizures. *Annals of Neurology*, *62*(2), 112–120.

Tapani, K. T., Vanhatalo, S., & Stevenson, N. J. (2019). Time-Varying EEG Correlations Improve Automated Neonatal Seizure Detection. *International Journal of Neural Systems*, *29*(04), 1850030.

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

Temko, A., Marnane, W., Boylan, G., & Lightbody, G. (2015). Clinical implementation of a neonatal seizure detection algorithm. *Decision Support Systems*, *70*, 86–96.

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

Uria-Avellanal, C., Marlow, N., & Rennie, J. M. (2013). Outcome following neonatal seizures. *Seminars in Fetal and Neonatal Medicine*, *18*(4), 224–232.

van de Sande, K. E. A., Uijlings, J. R. R., Gevers, T., & Smeulders, A. W. M. (2011). Segmentation as selective search for object recognition. *2011 International Conference on Computer Vision*, 1879–1886.

Widrow, B., & Hoff, M. (1960). Adaptive switching circuits. *IRE WESCON Convention Record*, *4*, 96–104.

Wolpaw, J. R., Birbaumer, N., Heetderks, W. J., McFarland, D. J., Peckham, P. H., Schalk, G., Donchin, E., Quatrano, L. A., Robinson, C. J., & Vaughan, T. M. (2000). Brain-computer interface technology: A review of the first international meeting. *IEEE Transactions on Rehabilitation Engineering: A Publication of the IEEE Engineering in Medicine and Biology Society*, *8*(2), 164–173.

Yıldız, E. P., Tatlı, B., Ekici, B., Eraslan, E., Aydınlı, N., Çalışkan, M., & Özmen, M. (2012). Evaluation of Etiologic and Prognostic Factors in Neonatal Convulsions. *Pediatric Neurology*, *47*(3), 186–192.

Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, *2*(10), 719–731.

Chapter 2: Monitoring neonatal seizures

2.1 Introduction

The neonatal brain is a complex, vulnerable, and rapidly developing organ which is central to an infant's growth and overall quality of life. Seizures in the neonatal period represent a clinical emergency and are the most common neurological complication seen in the NICU; neonatal seizures require timely detection and intervention.

Continuous EEG monitoring represents the gold-standard for neonatal seizure detection, but the interpretation of this complex signal is difficult and time consuming (Glass & Shellhaas, 2019) particularly for non-neurophysiologists. The need for automated decision support when monitoring neonatal EEG has been highlighted by many neonatal neurophysiologists (Boylan et al., 2010; Ramantani et al., 2019; Rennie et al., 2019). Algorithms with the ability to automatically detect neonatal seizures would assist clinical staff in the laborious and resource consuming task of annotating continuous EEG traces.

This chapter highlights the clinical importance and the challenges of brain monitoring for neonatal seizure detection. This chapter also outlines the datasets of neonatal EEG signals that were used in experiments in this thesis.

2.2 The structure and development of the neonatal brain

The first few hours, days, and weeks of human life is a time of immense growth and development. The infant develops quickly and adapts to life outside the amniotic sac. The 28 days immediately after birth are referred to as the neonatal period. In some cases, either because of suspected illness, premature birth or for a variety of other reasons, infants are admitted to the NICU after birth. The role of the NICU is to provide an environment which can support the critical needs of the sick or premature infant (Gibbins et al., 2008). An array of physiological signals are monitored for each infant. The information gathered is customised for each patient depending on their clinical needs.

One important differentiating factor between groups of infants in the NICU is their gestational age (GA) at birth. During pregnancy the GA is measured in weeks from the mother's last menstrual period (Behrman et al., 2007). An infant born between 37-42 weeks GA is

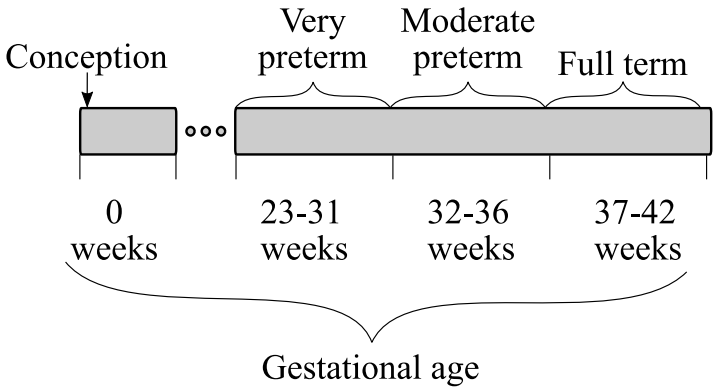


Figure 2.1 A representation of gestational age groups with respect to conception. Infants born before 37 weeks gestation fall into the category of preterm. A knowledge of GA can help to inform the correct treatment and care to provide to newborn infants.

considered to be a full term infant. All infants born before 37 weeks GA are recorded as preterm births. There is variability in the needs and risks associated with different GA groups. Within the window of preterm birth, infants are further categorized into extremely/very preterm and moderately/late preterm. In Cork University Maternity Hospital (CUMH), of approximately 8,500 births annually around 100-120 are 'very' or 'extremely' preterm category (Lloyd et al., 2017). The different GA groups are represented in Figure 2.1 with respect to the time of conception.

Recent years have seen the incorporation of rich and diverse ways of monitoring the neonatal brain. Often neuroimaging techniques are required to diagnose the underlying neurological problems in neonates. Cranial ultrasound (cUS) is a neuroimaging technique which uses ultrasound waves to map the structure of the brain; it is most effectively utilised in the neonatal setting due to the characteristics of the neonatal skull. In the neonatal period there are soft spots between the bones of the skull known as fontanelle, these spaces close over the first 12-18 months after birth. These open fontanelles provide unique access to the neonatal brain for the cUS allowing the infant brain to be imaged using cUS. The quality of images obtained through cUS varies depending on the operator's skill which limits the effectiveness of cUS in detecting brain injuries (Ramantani et al., 2019; Weeke et al., 2015b). The use of magnetic resonance imaging (MRI) in the neonatal setting has increased in recent years. MRI is more accurate in detecting underlying cranial abnormalities in neonates, but it usually requires the transfer of patients from the NICU to an MRI unit (Weeke et al., 2015a).

The use of neuroimaging techniques gives a snapshot of the neonatal brain structure and has high spatial resolution but low temporal resolution. Continuous monitoring techniques offer

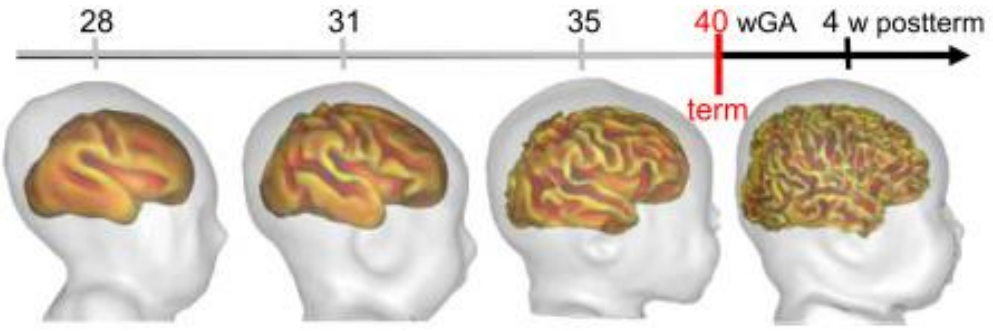


Figure 2.2 Degree of cortical folding of the neonatal brain relative to gestational age (Dehaene-Lambertz & Spelke, 2015). The gestational period is an important time of development for the vital organs, the brain is no exception to this. This visualisation of cortical folding shows the difference in brain development between an infant that is born prematurely and a full term infant.

spatial and high temporal resolution, they can provide a unique insight into the real-time function of the newborn brain. The application of near-infrared spectroscopy (NIRS) is a technique used to monitor cerebral oxygenation, but it is not currently in standard clinical use (Greisen, 2006). It has been theorised for many years that signal processing and machine learning can assist in the neural monitoring of these at-risk populations in the NICU. These innovations increase the potential benefits of domain specific engineering research in the NICU.

The brain is a centralised command hub for the entire body, exhibiting control over all organs and physiological systems. The neonatal brain is especially vulnerable to brain injury; any damage in early life can cause long-lasting effects. Electrical signals generated in the brain are an important window into understanding the status of the entire human body (Sanei & Chambers, 2013). In infants born prematurely, the central nervous system may still be developing and the cortical folding that happens during gestation may have been interrupted (Pavlidis et al., 2017; Ramantani et al., 2019). Figure 2.2 shows a representation of the cortical folding that takes place in the brain during gestation and in the neonatal period.

The brain is generally composed of nerve cells (neurons) and glia cells. Neurons respond to electrical stimuli and are capable of transmitting signals over long distances. The glia cells are located between neurons providing structural support (Sanei & Chambers, 2013). Each neuron consists of axons, dendrites, and the cell body. Neurons are massively inter-connected; each neuron is connected to approximately 10,000 other neurons. These connections are called synapses and these synaptic connections can be junctions between axons and dendrites, or between dendrites and dendrites. Figure 2.3 shows important the features of a neuron. The axon

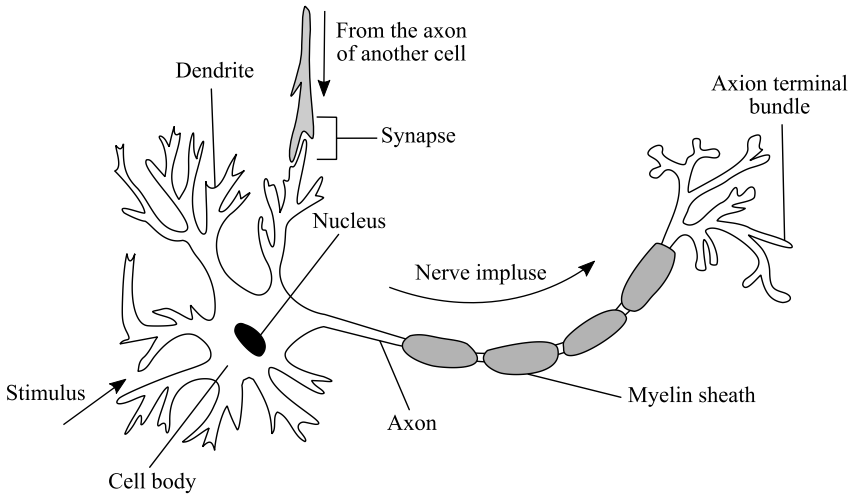


Figure 2.3 A representation of the structure of a neuron. The flow of action potential from the cell body, along the axon is indicated with an arrow. The brain is composed of a vastly interconnected network of neurons. This image was adapted from the web (Muskopf, 2020).

is a long, thin, cylinder which transmits electrical impulses. Axons are wrapped in insulating myelin sheaths which are regularly spaced out along the length of the axon; this increases the velocity of the transmitted signal (Kandel et al., 2013). Dendrites branch out from the cell body and receive electrical impulses from other neurons.

The signal that is propagated along the axon is called the action potential. The action potential is initiated in the cell body, where a stimulus is received. If the stimulus is above a certain threshold then the action potential will result in a temporary change in potential which is transmitted along the axon (Sanei & Chambers, 2013). The action potential appears as a spike and in most neurons, it lasts between 5 and 10 milliseconds. The action potential has an 'all-or-nothing' response, where the signal evoked by a large stimulus which far exceeds the threshold and a stimulus which is just marginally above the initiation threshold will both have the same size and shape (Kandel et al., 2013).

2.3 Neonatal electroencephalogram acquisition and monitoring

The electroencephalogram (EEG) is a measure of the electrical activity in the cerebral cortex. The electrical potentials are measured using electrodes placed on the surface of the scalp. The electrical signals recorded are a summation of the individual synaptic potentials in a specific region of the brain. Only large populations of active neurons generate enough potential to be recorded at the scalp due to signal attenuation in the skull and soft tissue.

Neonatal EEG signals are typically in the order of μV. Healthy background EEG signals are considered to be a $1/f$ process, where the power spectral density is inversely proportional to

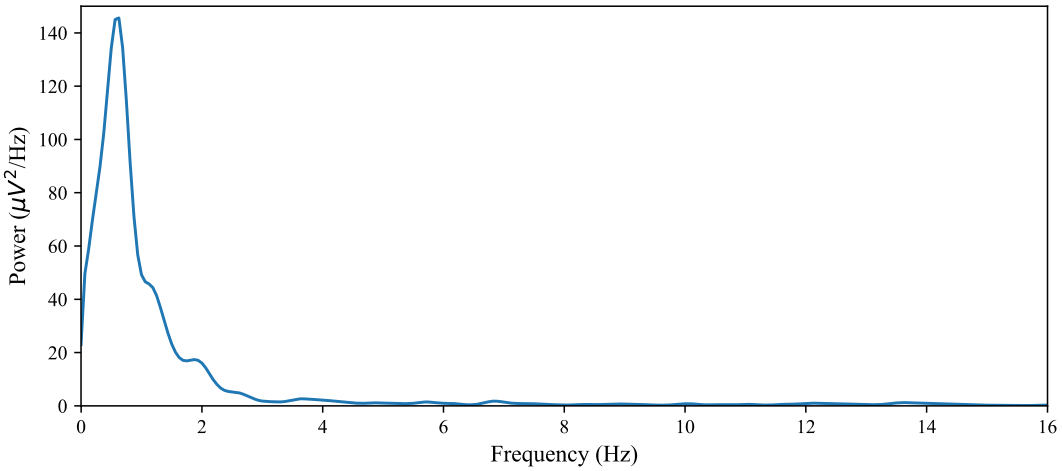


Figure 2.4 A plot of the power spectral density (PSD) of one infant shows that the signal is comparable to a 1/f process, alternatively this can be referred to as pink noise. This 60-second segment is taken from a period of EEG activity with no seizure events. The effect of the 0.5Hz high pass filter is also visible in the PSD.

the frequency (Rankine et al., 2007). Figure 2.4 shows a power spectral density plot from a 60s sample of neonatal EEG, the power spectrum in this figure approximates $1/f$ behaviour. The values on the y-axis are measured in µV2/Hz which is the signal power in each 1Hz band.

Neonatal EEG signals are characterised into different frequency bands, these different bands are used in clinical settings to determine the biological significance of rhythmic activity, see Table 2.1. When term neonates are awake their EEG usually consists of delta and theta activity. Quiet sleep in term infants is characterised by periods of delta and theta activity intermingled with faster periods of alpha and beta activity (Tsuchida et al., 2013). Analysis of the frequencies present in EEG at particular point in time provides clinical staff with insights into brain health and development.

Discontinuous features are also utilised by clinical staff to monitor neonatal brain health. The burst suppression pattern (which is an abnormal pattern) is characterised by a sudden burst of

Table 2.1 Frequency bands in neonatal EEG. Classifying EEG activity by the dominant frequencies present helps clinical staff to detect brain abnormalities and to detect different neural states e.g. sleep, seizure, and alertness.

Band	Range (Hz)
δ	0.5-4
θ	4-7.5
α	8-13
β	14-26
γ	>30

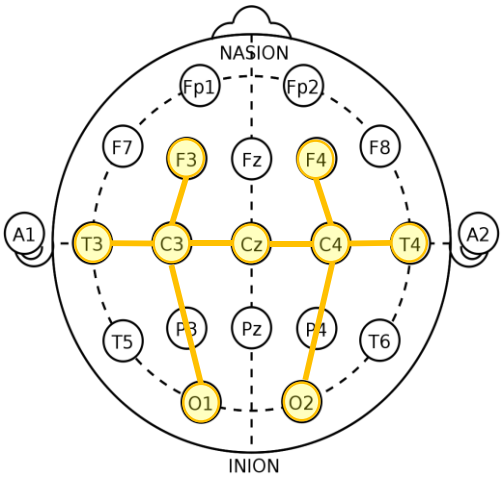


Figure 2.5 The 10-20 placement system modified for neonates. The yellow electrodes which are interconnected by yellow lines show one possible map of bipolar electrode pairs.

high energy, high frequency activity followed by a period of low voltage (Alix et al., 2017). Preterm EEG also exhibits discontinuous patterns, the characteristics of these discontinuous patterns are closely related to the age of the preterm infant. Normal term EEG is continuous. Asymmetry occurs when the EEG waveforms are unequal for the same areas on the opposite hemispheres of the brain. Usually, the EEG is synchronous over both brain hemispheres, however, occasional asymmetry is considered normal, particularly in preterm infants. The presence of long periods of sharp waves and spikes, which are substantially different from background EEG, can also indicate abnormal brain activity.

EEG is acquired using non-invasive electrodes, which are placed at specific locations on the scalp. The international 10-20 placement system is used, but the electrode placement is modified to account for the smaller size of the neonatal head. Figure 2.5 shows the electrodes which are utilised in the 10-20 placement system modified for neonates. In clinical practice, neonatal EEG is usually recorded with metal electrodes which sit on the surface of the infant's scalp. Due to the small amplitudes of EEG signals these surface electrodes can become over saturated with noise, so a skin preparation process is used to improve the electrical conductivity of the skin. This preparation process typically includes the application of an abrasive gel, to remove the outermost layer of skin cells. Recent research has compared the performance of dry electrodes, electrodes which require no skin preparation, with positive results, but these electrodes are not standard in current clinical practise (Di Flumeri et al., 2019; O'Sullivan et al., 2019).

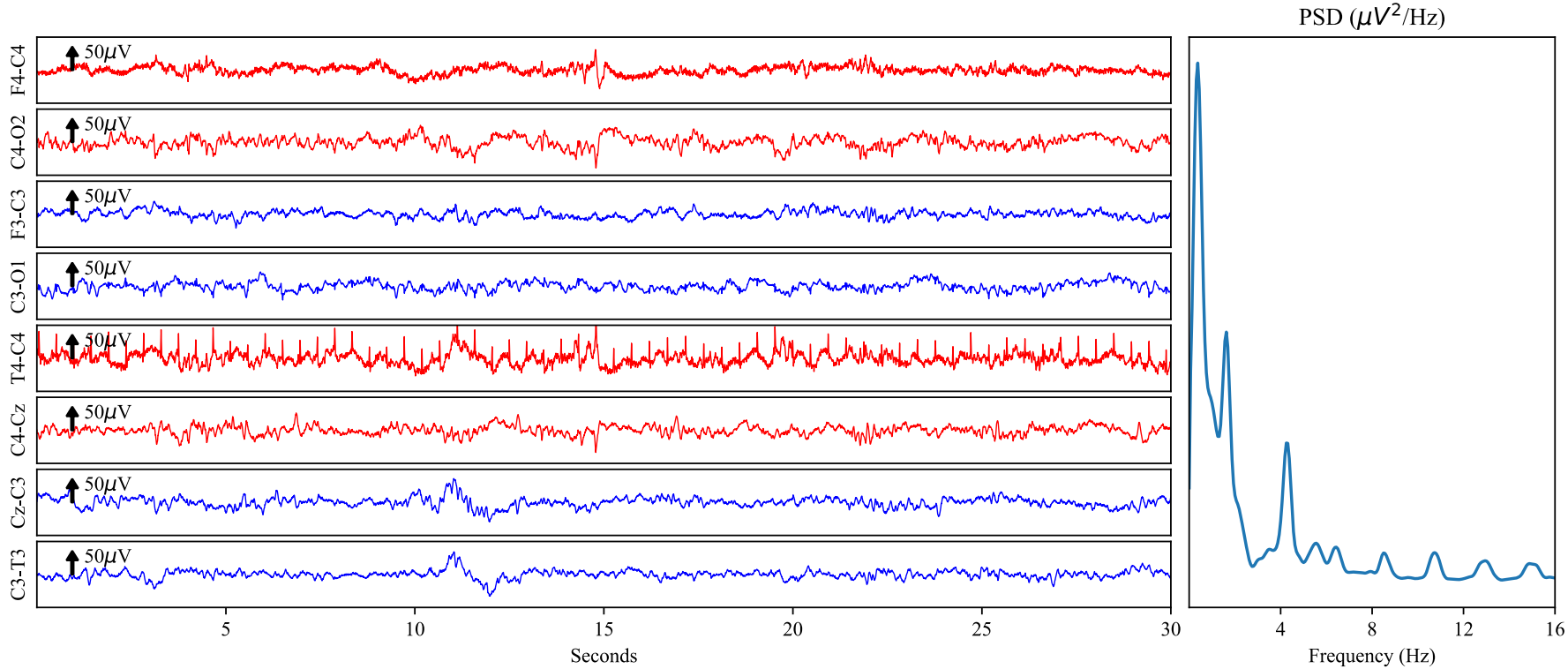


Figure 2.6 EEG with ECG artifact and the power spectral density plot of T4-C4 channel. The repetitive spike like pattern in the T4-C4 channel is most likely caused by an ECG artifact, the PSD plot shows increased signal power near the 1.5Hz and 4.25Hz frequencies.

The EEG signal recorded by each electrode is amplified and processed through an analogue-to-digital converter. This amplified digital signal is then displayed on a computer monitor and occasionally also saved in a patient's electronic medical record. EEG is typically digitised with a minimum sampling rate of at least 200Hz (Nuwer et al., 1998). EEG is recorded with respect to a single referential channel, where the recorded signal represents the difference between the electrode in question and a common referential electrode. In this work, EEG is interpreted as sets of bipolar channels, where each channel pair is displayed as the difference between two individual referential channels. See Figure 2.5 for a map of bipolar electrode pairs which utilises the 10-20 placement system modified for neonates.

EEG signals are susceptible to artifacts, these can be physiological or environmental in origin. Artifacts make EEG interpretation more complex; their presence can mask subtle signs of illness or they can mimic important diagnostic events. Physiological artifacts can be caused by movement, respiration, and pulse. Environmental noise can originate from the electrical mains and electrode detachment.

Figure 2.6 shows a segment of EEG where one channel has been corrupted with artefact from the electroencephalogram (ECG) signal; the repetitive spikes in channel T4-C4 are due to the rhythmic QRS electrical pattern generated by the infant's heart. The power spectral density of the T4-C4 channel shows elevated power at 1.5Hz and 4.25Hz, potentially due to the presence of ECG artifacts in the EEG.

2.4 Neonatal seizures

Seizures in newborns are an important indicator of underlying brain injury, often they represent the first clear sign that there is some neurological abnormality. Seizure incidence in the neonatal period is higher than at any other stage of life (Ramantani et al., 2019). Seizures occur in 1-3 of 1000 live births (Glass et al., 2016; Lanska et al., 1995; Ronen et al., 2007; Saliba et al., 2001), higher rates are reported in preterm births (Kohelet et al., 2004). This makes seizures the most common neurological emergency in the NICU.

A seizure is defined as a paroxysmal alteration in neurologic function, the neurologic functions involved can be behavioural, motor, or automatic (Volpe, 2008). The clinical manifestation in the neonatal period varies from seizures in adults and children. The manifestation of neonatal seizures varies widely, and clinical indicators are often extremely subtle. The differences in seizure manifestation in the neonatal period are due to developmental, structural, and neurochemical differences (Ramantani et al., 2019). Studies have shown that up to two thirds

of seizures in the NICU are subclinical, meaning they have no visible signs (Murray et al., 2008). Under typical NICU conditions, where only visual inspection is available to clinical staff, less than 10% of neonatal seizures are detected (Murray et al., 2008).

The most common cause of seizures in full term infants is hypoxic-ischemic encephalopathy (HIE) and the most common cause in preterm infants is interventricular haemorrhage (IVH), although a wide variety of other etiologies can lead to neonatal seizures (Ramantani et al., 2019). Outcomes after a diagnosis of neonatal seizures are worse in preterm neonates, but the survival rate in the vulnerable preterm population is increasing (Pisani & Spagnoli, 2018). Recent evidence suggests that the seizure events themselves cause additional damage to the brain after the initial insult; this has yet to be definitively proven, but it certainly motivates the prompt treatment of neonatal seizure events (Kharoshankaya et al., 2016; Sands & McDonough, 2016).

2.4.1 Methods to detect neonatal seizure events

The detection of seizures in the NICU is an important step in the diagnosis of brain injury. Timely treatment of seizures is important, as seizure events cannot be retrospectively treated. Seizures pose a significant diagnostic challenge for NICU staff.

The most accessible method of detecting seizures is through visual inspection of the neonate. No special equipment or preparations are required. Around a third of neonatal seizures have physical signs and can be visually diagnosed if the medical staff know which movements to look for and have the time to frequently inspect the infant (Murray et al., 2008). The clinical

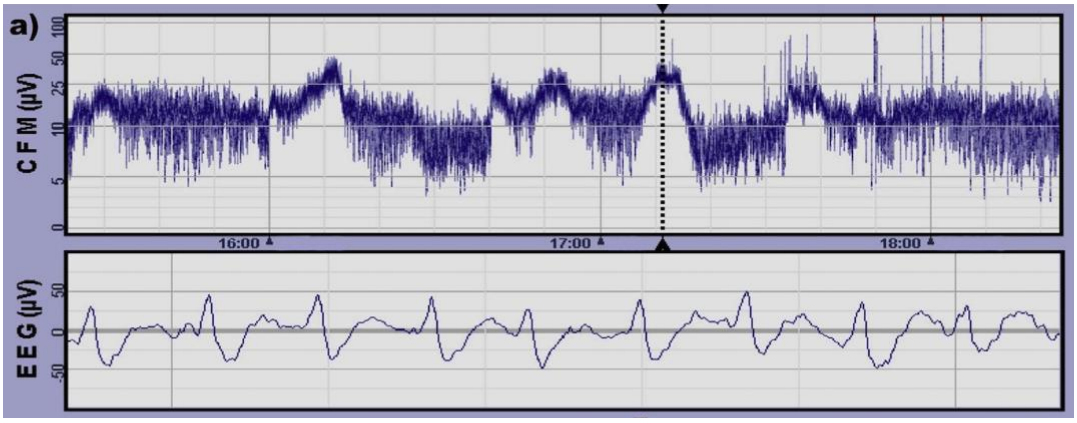


Figure 2.7 The upper plot shows the aEEG trace over three hours of neonatal EEG. The lower plot highlights 6 seconds of single channel EEG which is centred around the time point marked by a vertical dashed line on the aEEG. The time indicated by the vertical dashed line is an example seizure event taken from Hart et al. (Hart et al., 2017). The repetitive spike pattern in the raw EEG signal is a sign of neonatal seizure. The sections of aEEG where the bandwidth narrows, and the voltage increases are indicative of seizure events.

signs that characterise neonatal seizures are often subtle movements, which are easily overlooked by busy staff.

EEG is the most reliable method of detecting all neonatal seizures, but it requires specialists to interpret the complex EEG signals (Ramantani et al., 2019, Temko & Lightbody, 2016). Trained neurophysiologists scroll through multiple pages of EEG readings to check if the pattern which is suspected of being a seizure event can be confirmed. This is the gold-standard for seizure detection in neonates but continuous access to EEG monitoring equipment and expertise is not available in many NICUs.

Amplitude integrated EEG (aEEG) is a limited version of EEG which simplifies the signal representation. aEEG is also known as cerebral function monitoring, an example of a seizure event captured in aEEG and raw temporal EEG is shown in Figure 2.7. The advantage of this representation is that it can be interpreted by less specialised staff and many hours of EEG activity can be viewed in one plot. The increased accessibility of aEEG comes at the cost of decreased sensitivity (Temko & Lightbody, 2016). aEEG is limited to just one or two channels;

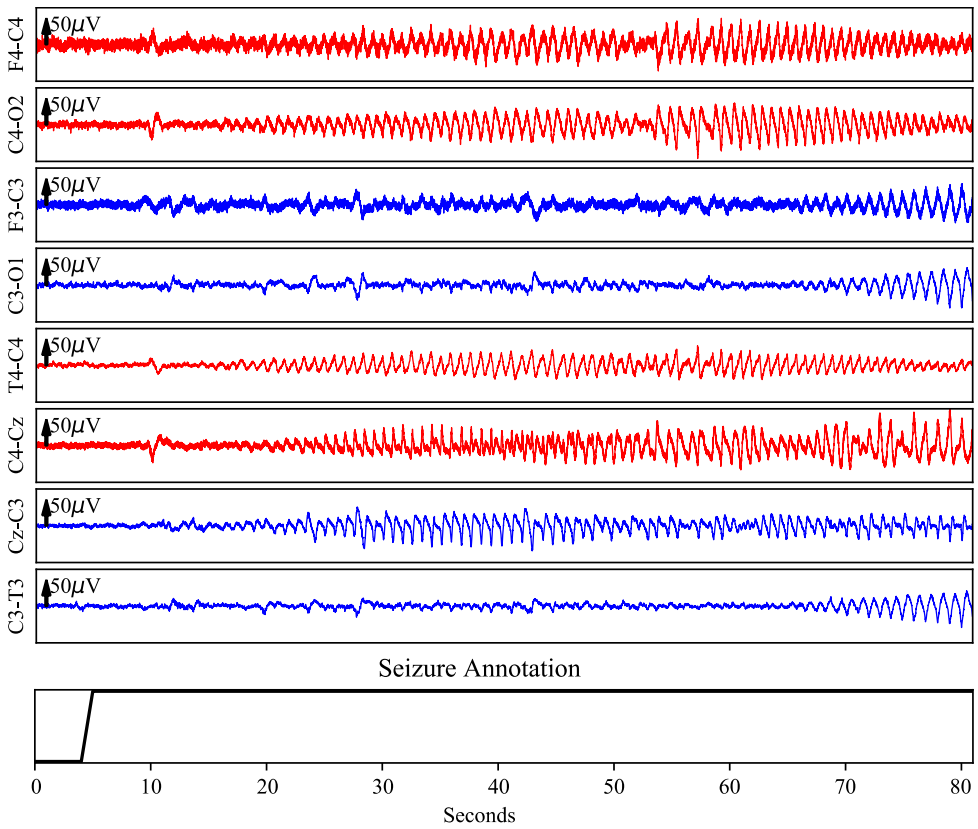


Figure 2.8 A seizure migrating across channels. The red and blue signals are representative of the different hemispheres of the brain (blue – left, red – right). The repetitive seizure pattern starts on the right side of the brain and migrates over time to the left side of the brain.

electrode pairs near the centre of the infant's scalp are typically selected. The raw EEG signal is filtered, rectified, processed, and displayed with a semi-logarithmic amplitude on a time-compressed scale (Glass et al., 2013). Short duration seizures and seizures of low amplitude may be filtered out of the aEEG trace, this reduces the sensitivity of aEEG as a seizure detection tool.

2.4.2 The manifestation of neonatal seizures in EEG

Neonatal seizures in EEG are defined as a repetitive, evolving pattern with a voltage greater than 2μV which lasts for longer than 10 seconds (Ramantani et al., 2019). The background of neonatal EEG is characterised as being pseudo-random with no repeating patterns. In contrast to this baseline EEG behaviour, seizures represent an ordered, rhythmic, and structured deviation from the non-seizure EEG. Seizures are also described as evolving, this evolution can occur in the dominant seizure frequencies, the power of the seizure, and in the spatial location of the seizure (Tsuchida et al., 2013).

Figure 2.8 presents a seizure event which migrates across channels. The seizure starts in the right hemisphere of the brain (red channels) and migrates to the left hemisphere (blue

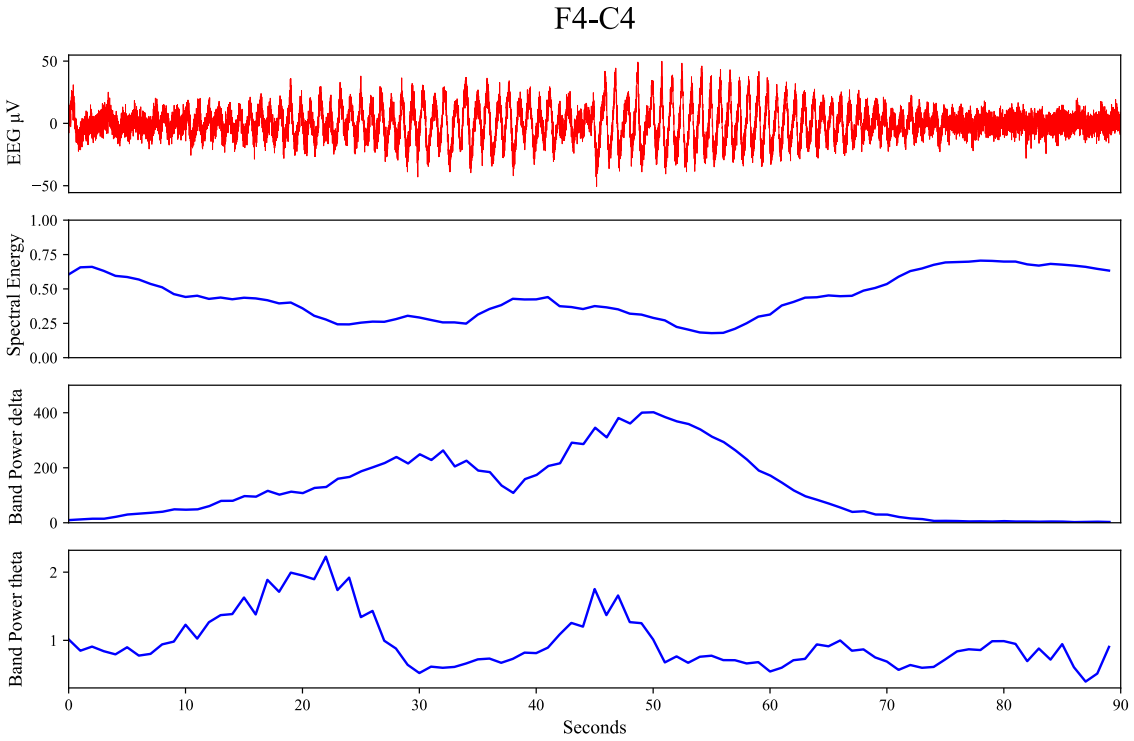


Figure 2.9 A seizure event with features plotted. The seizure increases in power in both the theta and delta frequency bands over time. The entropy of the signal reduces as the EEG pattern becomes more repetitive.

channels). This seizure is a clear deviation from the background pattern, manifesting as a repetitive high-energy event.

Figure 2.9 shows the F4-C4 channel from the example in Figure 2.8 plotted alongside features extracted from the waveform. The spectral entropy, which measures uncertainty in the normalised power in frequency bands of a signal, shows that the signal becomes more ordered and less random during the seizure event. Most neonatal seizures are in the δ and θ frequency bands, in Figure 2.9 the power in these two frequency sub-bands increases during the seizure event.

2.4.3 Challenges associated with neonatal seizure detection

Detecting and treating neonatal seizures poses a considerable clinical challenge. Under-diagnosis and over-diagnosis are both common problems due to the uncoupling between clinical movements and electrographic seizures (Rennie & Boylan, 2003). It is not possible to have neurophysiological experts available at all times in the NICU to interpret EEG signals. The development of clinically supportive algorithms could improve seizure treatment, reducing the impact of neonatal seizures on the developing brain.

A major challenge to consider when developing seizure detection algorithms is inter-patient generalisation. Each infant in the NICU will have different baseline EEG characteristics, trained seizure detection algorithms should be able to account for this variability. The development of automated seizure detection algorithms is an ever-evolving field. Early automated seizure detection routines relied on heuristic rules and thresholds. The clinical characteristics of neonatal seizures prompted the search for features that characterise the repetitiveness, order and predictability of ictal EEG, and the subsequent application of a designed ruleset or classifier to these features (Celka & Colditz, 2002; Gotman et al., 1997; Liu et al., 1992). These early classifiers struggled with the diverse nature of neonatal seizures. Each algorithm could detect certain seizure patterns but failed when the morphology did not match the specific patterns they were tuned to detect (Faul et al., 2005).

The development of more complex and data-driven algorithms did improve the robustness and generalisability of classifiers, but the presence of artifacts still poses a problem for classifiers (Tapani et al., 2019; Temko et al., 2013). Modern data-driven machine learning algorithms are also limited by the amount of data and the quality of the seizure labels in their training dataset. The collection of a large dataset of clean and clearly labelled neonatal EEG files poses a

considerable challenge; the lack of a large well-annotated dataset makes the algorithm development process more challenging.

2.5 Clinical annotations of neonatal seizure events

Clinical staff in NICUs are focused on detecting and treating seizure events. They consider the time and the duration of these events to be important when deciding treatment options, but the location and the spread across channels of these events is secondary and is not routinely recorded. Seizures in the neonatal brain are often localised in nature, meaning they can be present in just a single channel or a couple of EEG channels. This focal nature of neonatal seizures makes the task of creating clean datasets of annotated seizure events challenging.

2.5.1 Strong and weak labels

Seizures can be annotated in different ways. First, the infants can be annotated as having experienced seizures; this differentiates them from infants who did not have seizures. While these labels are clinically valuable, they provide no spatial or temporal information about the seizures and are not suitable for developing automatic seizure detection systems. Second, seizures can be localised in time with overall annotations, whereby the start and end time for each seizure event is recorded for each file giving precise temporal resolution to each seizure event, this is a weak label. Figure 2.10 (a) shows weak annotations for a seizure, where the start and end times of the seizure are indicated, but the channels involved in the seizure event are not specified. Propagating these overall annotations to each EEG channel will introduce a lot of noise, since seizures can be localised to a single channel and even migrate from one channel to another during the seizure event. These overall annotations are sufficient for evaluation of algorithmic performance but are not enough for the extraction of seizure signal signatures to train a classifier. In this work, we refer to these annotations as weak annotations.

Thirdly, in addition to overall annotations, seizures can be localised in space, whereby each seizure (or a part of it) has not only its start and end time but also its spatial location specified as shown in Figure 2.10 (b). We will refer to these as strong labels. There is a large cost associated with the acquisition of strong labels as it requires a lot of extra time and effort from clinical staff. In the datasets of neonatal seizures used in this work a small minority of the seizure events have associated strong labels, due to the large cost associated with acquiring strong labels. Due to the migratory nature of seizures, clearly defining the edges of a seizure event in the representative channels is a challenge for experts, instead strong seizure labels are usually given to subsections taken from the middle of seizure events.

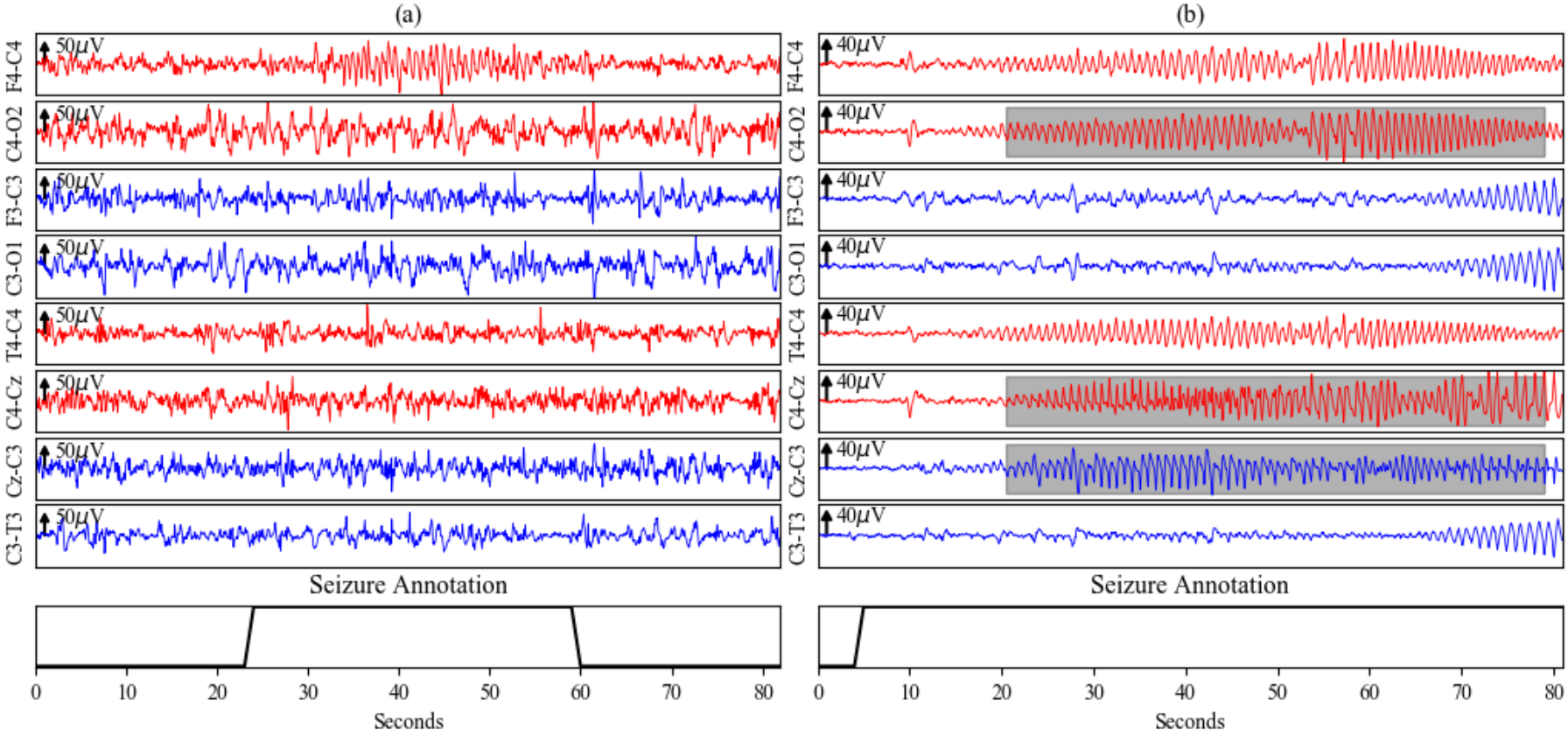


Figure 2.10 (a) shows an example of a seizure event with a weak label. A seizure is observable only in the F4-C4 channel, which cannot be derived from the weak label (shown below the EEG trace).(b) shows an example of a seizure event with a weak label (shown below the EEG trace), some parts of the seizure event also have strong labels. The strong labels in (b) localise the seizure activity (shown as grey shading). However, this strong label does not imply that these are the only channels where the seizure is present, but these are the only channels where a strong label is available

2.5.2 Inter-observer agreement

EEG interpretation can be subjective, even the most experienced experts may disagree when analysing EEG from the same patient. The level of agreement between EEG reviewers can be quantified using the inter-observer agreement (IOA) metric. Typically, the IOA is calculated using both temporal and event-based metrics. Short duration seizures usually result in higher levels of disagreement between experts. The median agreement for seizure burden has been reported as 83%, this implies that any automated seizure detection algorithm would be expected to have a similar IOA when compared to the annotations of a trained expert (Stevenson et al., 2015).

2.6 Datasets of neonatal EEG signals

2.6.1 Term Dataset I – Cork

A dataset of EEG signals recorded in the NICU of CUMH, Ireland was used in this work. The Cork dataset includes recordings from 18 full term newborn infants, between 39 and 42 weeks GA, all of whom experienced seizures because of brain injury associated with HIE. The Cork dataset totals over 834 hours in duration and comprises of 1389 seizure events. More details about the seizure events can be seen in Table 2.2. All files contain eight channels of EEG, recorded using the 10-20 placement system modified for neonates. A bipolar montage of the

Table 2.2 The Cork dataset of full term neonatal EEGs with seizure annotations.

Infant Number	Record length (h)	Seizure events	Mean seizure duration	Min seizure duration	Max seizure duration
1	29.66	17	1’29’’	17’’	3’54’’
2	24.74	3	6’10’’	55’’	11’9’’
3	31.43	209	1’50’’	11’’	10’43’’
4	47.54	84	1’37’’	32’’	9’58’’
5	56.78	63	6’32’’	20’’	34’10’’
6	19.24	46	1’7’’	15’’	4’17’’
7	60.75	99	1’31’’	14’’	10’20’’
8	49.55	17	5’55’’	29’’	19’14’’
9	67.76	201	4’59’’	13’’	37’6’’
10	59.79	41	4’51’’	13’’	34’46’’
11	22.21	43	2’27’’	17’’	7’36’’
12	54.36	150	1’35’’	15’’	10’8’’
13	51.68	60	3’26’’	19’’	16’56’’
14	22.79	21	8’12’’	22’’	39’3’’
15	66.02	121	1’30’’	10’’	7’8’’
16	76.39	189	5’4’’	26’’	34’37’’
17	30.7	21	5’30’’	27’’	23’16’’
18	63.02	4	9’33’’	7’19’’	13’22’’
Total	834.41	1389			

following eight channels was used: F4-C4, C4-O2, F3-C3, C3-O1, T4-C4, C4-Cz, Cz-C3, C3-T3. Raw multi-channel EEG data were collected using the Carefusion NicOne video EEG monitor at a sampling rate of 256Hz.

Two independent neonatal electro-encephalographers annotated all seizure events in this dataset. The agreement of these experts was taken as the ground truth and any disagreements were discussed until consensus was reached by both experts. The experts had the EEG waveforms and a synchronised video recording of the cot available to them while annotating the seizure events. No data pre-selection or artifact rejection through engineered rules was performed, to avoid human-induced biases. The artifacts which are normally present in long-term EEG will form part of the data-driven training process. The data are reflective of a real-world NICU environment where EEG signals would contain a variety of different artifacts. This study had full ethical approval from the Clinical Research Ethics Committee of the Cork Teaching Hospitals.

2.6.2 Term Dataset II – Helsinki

A publicly available dataset of annotated neonatal EEG waveforms collected at Helsinki University Hospital was also used as a test dataset in this work (Stevenson et al., 2019). This dataset contains 18-channel EEG segments from 79 full term infants; signals are sampled at 256Hz. The dataset is not composed of long continuous unedited recordings but rather contains 1-2h excerpts per infant. As a result, in contrast to the Cork dataset, this dataset is larger in the number of infants (79 vs 18 in Cork) but much smaller in both the overall duration (112 hours of recordings vs 834 hours in Cork), and in the number of seizure events (342 vs 1389 in Cork). The seizure annotations were defined based on the consensus of three EEG experts. Out of 79 infants, 39 experienced seizures.

2.6.3 Preterm Dataset I – Parma

The first preterm dataset was collected in Parma University Hospital, Italy. Preterm infants with confirmed video-EEG seizures, were admitted to the NICU of Parma University Hospital. A sample dataset was then created with the following inclusion criteria: 1) infants born <32 weeks GA; 2) v-EEG-confirmed neonatal seizures; 3) seizures onset <36 weeks GA. Ethical approval for the collection and analysis of the data was granted by the Ethics Committee of Parma. Written informed parental consent was obtained. Recordings were started on average 13 days (range 1-47 days) after birth.

Table 2.3: The Parma dataset of preterm EEGs with seizure annotations.

Age (Weeks)	Record length (hh:mm:ss)	Seizure events	Seizure event durations	
			> 1-min	< 1-min
24	4:49:01	26	14	12
24	1:29:55	2	2	0
25	0:37:23	12	6	6
26	1:48:40	8	0	8
26	0:32:48	1	1	0
27	1:02:02	3	0	3
27	2:00:16	12	9	3
28	1:54:46	13	1	12
28	0:57:23	1	1	0
29	1:31:36	2	0	2
29	1:01:46	1	1	0
30	1:10:04	5	2	3
30	0:38:54	1	1	0
30	1:10:35	15	8	7
31	1:16:29	4	4	0
32	0:51:38	21	17	4
32	0:31:03	1	1	0
Total	23:24:19	128	68	60

Depending on the infant's head size, electrodes were applied according to the international 10-20 system modified for infants using 14 recording channels. Recordings continued until a complete cycle of wakefulness, quiet, and active sleep was obtained. When state changes were not clearly distinguishable, the recording continued for at least 60 min.

Table 2.3 details the Parma dataset. Whenever an infant had several recordings in the same day or serial follow-up recordings, only the files with at least one seizure event were selected for inclusion in the dataset. The dataset mainly comprised short EEG recordings lasting an average of 1 hour and 19 minutes (range: 0.5-4.5 hours). Half of all seizures in this dataset were less than 1-minute duration (short seizures).

2.6.4 Preterm Dataset II – Cork

The second preterm dataset was recorded in CUMH, Ireland. This is a dataset of long-recordings from infants less than 32 weeks of gestation (Lloyd et al., 2017). A modified 10-20 neonatal electrode montage was used; 8 channels were selected for analysis: F4-C4; C4-O2; F3-C3; C3-O1; T4-C4; C4-Cz; Cz-C3; C3-T3. All seizure events in this dataset are annotated, but no information about the seizure channel locations is available. Ethical approval for the collection and analysis of the data was granted by the Clinical Research Ethics Committee of the Cork Teaching Hospitals, Ireland. Written informed parental consent was obtained for each study.

Table 2.4: The Cork dataset of preterm EEGs with seizure annotations.

Age (Weeks)	Record Length (hours)	Seizure events	Seizure durations		Seizure statistics		
			>1min	<1min	Mean duration	Min duration	Max duration
23	24.00	0	-	-	-	-	-
24	48.02	84	24	60	48''	12''	5'46''
24	24.02	0	-	-	-	-	-
25	58.80	6	0	6	25''	20''	29''
25	99.56	49	38	11	2'44''	29''	5'54''
25	12.02	0	-	-	-	-	-
26	12.01	0	-	-	-	-	-
26	24.00	0	-	-	-	-	-
26	67.10	1	0	1	41''	41''	41''
26	14.76	0	-	-	-	-	-
29	48.49	15	7	8	51''	33''	1'20''
29	12.15	0	-	-	-	-	-
29	12.09	0	-	-	-	-	-
30	24.06	0	-	-	-	-	-
31	24.03	0	-	-	-	-	-
31	70.13	149	94	55	1'26''	15''	22'32''
Total	575.24	304	163	141			

Table 2.4 presents the details of the Cork dataset. The dataset contains 6 seizure infants and 10 non-seizure infants. As the preterm EEG evolves depending on gestational age (GA), three groups were created from this dataset: Group 1 - GA < 26 weeks, Group 2 - 26 weeks ≤ GA ≤ 29 weeks, Group 3 - GA > 29 weeks. Around 50% of all seizures in this dataset were short seizures of less than 1-minute duration (Lloyd et al., 2017). The continuous EEG recordings in this dataset were not edited to remove the large variety of artefacts commonly encountered in the real-world NICU environment. Therefore, this dataset is truly representative of the real-life situation in the NICU.

2.7 Conclusion

Detecting neonatal seizures is a clinically import task, which continues to pose a challenge due to the resource constraints on clinical expertise and the complexity and variety of neonatal EEG signals. Many recent publications in the area of neonatal seizure detection call for the development of automated seizure detection algorithms, which can assist with this diagnostic challenge (Ramantani et al., 2019; Rennie et al., 2019). The task of developing automated methods for detecting seizures in neonatal EEG is complicated by the presence of artifacts in neonatal EEG, the variability between patients' baseline patterns, and the lack of available clinical data and corresponding strong clinical labels.

2.8 References

Alix, J. J. P., Ponnusamy, A., Pilling, E., & Hart, A. R. (2017). An introduction to neonatal EEG. *Paediatrics and Child Health*, *27*(3), 135–142.

Behrman, R. E., Butler, A. S., & Outcomes, I. of M. (US) C. on U. P. B. and A. H. (2007). *Measurement of Fetal and Infant Maturity*. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK11382/

Boylan, G. B., Burgoyne, L., Moore, C., O'Flaherty, B., & Rennie, J. M. (2010). An international survey of EEG use in the neonatal intensive care unit. *Acta Paediatrica*, *99*(8), 1150–1155.

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

Dehaene-Lambertz, G., & Spelke, E. S. (2015). The Infancy of the Human Brain. *Neuron*, 88(1), 93–109.

Di Flumeri, G., Aricò, P., Borghini, G., Sciaraffa, N., Di Florio, A., & Babiloni, F. (2019). The Dry Revolution: Evaluation of Three Different EEG Dry Electrode Types in Terms of Signal Spectral Features, Mental States Classification and Usability. *Sensors (Basel, Switzerland)*, *19*(6).

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

Gibbins, S., Hoath, S. B., Coughlin, M., Gibbins, A., & Franck, L. (2008). The universe of developmental care: A new conceptual model for application in the neonatal intensive care unit. *Advances in Neonatal Care: Official Journal of the National Association of Neonatal Nurses*, *8*(3), 141–147.

Glass, H. C., & Shellhaas, R. A. (2019). Acute Symptomatic Seizures in Neonates. *Seminars in Pediatric Neurology*, *32*, 100768.

Glass, H. C., Shellhaas, R. A., Wusthoff, C. J., Chang, T., Abend, N. S., Chu, C. J., Cilio, M. R., Glidden, D. V., Bonifacio, S. L., Massey, S., Tsuchida, T. N., Silverstein, F. S., Soul, J. S., Bergin, A. M., Dlugos, D., Ferriero, D. M., & Staley, K. (2016). Contemporary Profile of Seizures in Neonates: A Prospective Cohort Study. *The Journal of Pediatrics*, *174*, 98-103.e1.

Glass, H. C., Wusthoff, C. J., & Shellhaas, R. A. (2013). Amplitude-Integrated Electroencephalography: The Child Neurologist's Perspective. *Journal of Child Neurology*, *28*(10), 1342–1350.

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

Greisen, G. (2006). Is near-infrared spectroscopy living up to its promises? *Seminars in Fetal and Neonatal Medicine*, *11*(6), 498–502.

Hart, A. R., Ponnusamy, A., Pilling, E., & Alix, J. J. P. (2017). Neonatal cerebral function monitoring – understanding the amplitude integrated EEG. *Paediatrics and Child Health*, *27*(4), 187–195.

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A., & Hudspeth, A. (2013). *Principles of Neural Science, Fifth Editon*. Cenveo Publisher Services, Columbia, MD, USA.

Kharoshankaya, L., Stevenson, N. J., Livingstone, V., Murray, D. M., Murphy, B. P., Ahearne, C. E., & Boylan, G. B. (2016). Seizure burden and neurodevelopmental outcome in neonates with hypoxic–ischemic encephalopathy. *Developmental Medicine & Child Neurology*, *58*(12), 1242–1248.

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

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

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

Lloyd, R. O., O'Toole, J. M., Pavlidis, E., Filan, P. M., & Boylan, G. B. (2017). Electrographic Seizures during the Early Postnatal Period in Preterm Infants. *The Journal of Pediatrics*, *187*, 18-25.e2.

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

Muskopf, S. (2020). *The Anatomy of a Synapse*. Biology Corner. https://www.biologycorner.com/worksheets/synapse-coloring.html

Nuwer, M. R., Comi, G., Emerson, R., Fuglsang-Frederiksen, A., Guerit, J.-M., Hinrichs, H., Ikeda, A., Luccas, F. J. C., & Rappelsburger, P. (1998). IFCN standards for digital recording of clinical EEG. *Electroencephalography and Clinical Neurophysiology*, *106*(3), 259.

O'Sullivan, M., Temko, A., Bocchino, A., O'Mahony, C., Boylan, G., & Popovici, E. (2019). Analysis of a Low-Cost EEG Monitoring System and Dry Electrodes toward Clinical Use in the Neonatal ICU. *Sensors*, *19*(11), 2637.

Pavlidis, E., Lloyd, R. O., & Boylan, G. B. (2017). EEG - A Valuable Biomarker of Brain Injury in Preterm Infants. *Developmental Neuroscience*, *39*(1–4), 23–35.

Pisani, F., & Spagnoli, C. (2018). Acute symptomatic neonatal seizures in preterm neonates: Etiologies and treatments. *Seminars in Fetal and Neonatal Medicine*, *23*(3), 191–196.

Ramantani, G., Schmitt, B., Plecko, B., Pressler, R. M., Wohlrab, G., Klebermass-Schrehof, K., Hagmann, C., Pisani, F., & Boylan, G. B. (2019). Neonatal Seizures—Are We there Yet? *Neuropediatrics*, *50*(5), 280–293.

Rankine, L., Stevenson, N., Mesbah, M., & Boashash, B. (2007). A nonstationary model of newborn EEG. *IEEE Transactions on Bio-Medical Engineering*, *54*(1), 19–28.

Rennie, J. M., & Boylan, G. B. (2003). Neonatal seizures and their treatment. *Current Opinion in Neurology*, *16*(2), 177–181.

Rennie, J. M., Vries, L. S. de, Blennow, M., Foran, A., Shah, D. K., Livingstone, V., Huffelen, A. C. van, Mathieson, S. R., Pavlidis, E., Weeke, L. C., Toet, M. C., Finder, M., Pinnamaneni, R. M., Murray, D. M., Ryan, A. C., Marnane, W. P., & Boylan, G. B. (2019). Characterisation of neonatal seizures and their treatment using continuous EEG monitoring: A multicentre experience. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, *104*(5), F493–F501.

Ronen, G., Buckley, D., Penney, S., & Streiner, D. (2007). Long-term prognosis in children with neonatal seizures: A population-based study. *Neurology*, *69*(19), 1816–1822.

Saliba, R. M., Annegers, F. J., Waller, D. K., Tyson, J. E., & Mizrahi, E. M. (2001). Risk Factors for Neonatal Seizures: A Population-based Study, Harris County, Texas, 1992–1994. *American Journal of Epidemiology*, *154*(1), 14–20.

Sands, T. T., & McDonough, T. L. (2016). Recent Advances in Neonatal Seizures. *Current Neurology and Neuroscience Reports*, *16*(10), 92.

Sanei, S., & Chambers, J. A. (2013). *EEG Signal Processing*. John Wiley & Sons.

Stevenson, N. J., Clancy, R. R., Vanhatalo, S., Rosén, I., Rennie, J. M., & Boylan, G. B. (2015). Interobserver agreement for neonatal seizure detection using multichannel EEG. *Annals of Clinical and Translational Neurology*, *2*(11), 1002–1011.

Stevenson, N. J., Tapani, K., Lauronen, L., & Vanhatalo, S. (2019). A dataset of neonatal EEG recordings with seizure annotations. *Scientific Data*, *6*(1), 190039.

Tapani, K. T., Vanhatalo, S., & Stevenson, N. J. (2019). Time-Varying EEG Correlations Improve Automated Neonatal Seizure Detection. *International Journal of Neural Systems*, *29*(04), 1850030.

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

Temko, A., & Lightbody, G. (2016). Detecting Neonatal Seizures With Computer Algorithms. *Journal of Clinical Neurophysiology*, *33*(5), 394–402.

Tsuchida, T. N., Wusthoff, C. J., Shellhaas, R. A., Abend, N. S., Hahn, C. D., Sullivan, J. E., Nguyen, S., Weinstein, S., Scher, M. S., Riviello, J. J., & Clancy, R. R. (2013). American Clinical Neurophysiology Society Standardized EEG Terminology and Categorization for the Description of Continuous EEG Monitoring in Neonates: Report of the American Clinical Neurophysiology Society Critical Care Monitoring Committee. *Journal of Clinical Neurophysiology*, *30*(2), 161–173.

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

Weeke, L. C., Groenendaal, F., Toet, M. C., Benders, M. J. N. L., Nievelstein, R. A. J., Rooij, L. G. M. van, & Vries, L. S. de. (2015a). The aetiology of neonatal seizures and the diagnostic contribution of neonatal cerebral magnetic resonance imaging. *Developmental Medicine & Child Neurology*, *57*(3), 248–256.

Weeke, L. C., Rooij, L. G. M. van, Toet, M. C., Groenendaal, F., & Vries, L. S. de. (2015b). Neuroimaging in neonatal seizures. *Epileptic Disorders: International Epilepsy Journal with Videotape*, *17*(1), 1–11; quiz 11.

Chapter 3: Neonatal seizure detection methods and an introduction to deep learning

3.1 Introduction

The challenges related to detecting seizures in the NICU and the complexity and resources associated with manual analysis of long EEG recordings have motivated the development of automated seizure detection algorithms (Ramantani et al., 2019). The research into neonatal seizure detection algorithm development would not be possible without collaboration between medical experts and engineers. Over time the nature of these algorithms has shifted from purely feature-driven to more data-driven approaches (Temko & Lightbody, 2016). Since 2012, the advent of deep learning has paved the way for the development of purely data-driven neonatal seizure detection algorithms.

Section 3.2 of this chapter reviews previously developed neonatal seizure detection algorithms. In Section 3.3 two machine learning systems which are utilised as comparative baselines throughout the work are presented: the SVM Cork algorithm and the SVM Helsinki algorithm. The feature sets associated with these algorithms are outlined; feature sets are utilised to represent important signal characteristics which are known to facilitate the detection of seizures. Section 3.4 details the metrics and frameworks used to quantify performance of algorithms throughout this work. Section 3.5 outlines the deep learning methodology employed in this thesis. Specifically, the foundational algorithms for deep learning and how deep learning-based architectures could be applied in the neonatal seizure detection domain are explored.

3.2 Algorithms for neonatal seizure detection

3.2.1 Early works: mimicking the clinical approach

The first automated neonatal seizure detection algorithms sought to mimic the procedures followed by clinical staff when detecting seizure events. Other early proposed solutions were based on algorithms which were designed to detect seizures in adult EEG. The development of bespoke features became an important area of research for those trying to develop early neonatal seizure detection algorithms.

When a clinical neurophysiologist is analysing an EEG recording, they will search for repetitive, pseudo-periodic waveforms, which differ from the background EEG. The

waveforms could be smooth and oscillatory or manifest as a sharp, spike-like patterns. For an event to be classified as a seizure it must display a definite start, middle and end and have a duration of more than 10 seconds (Clancy & Legido, 1987). Analysing whether patterns morph over time is an important step to discriminate between seizures and rhythmic artifacts; seizures will often change in dominant frequency, amplitude and location over the duration of the seizure event (Patrizi et al., 2003). To analyse the temporal morphology of a suspected seizure, clinical neurophysiologists will scroll forwards and backwards in time to search for these indicators of signal evolution. Early algorithms utilised this clinical knowledge to create rule-based approaches to detect seizures (Gotman et al., 1997; Liu et al., 1992).

Liu et. al. developed a feature based neonatal seizure detection algorithm in 1992 which was designed to detect rhythmic behaviour in EEG (Liu et al., 1992). Epochs were characterised using time marks from successive peaks of an autocorrelation plot. Seizures should appear more rhythmic and have a more regular autocorrelation plot than non-seizure epochs; this hypothesis was informed by the clinical definition of a neonatal seizure.

In 1997 Gotman et. al. proposed a set of three neonatal seizure detection algorithms; each was designed to detect a different type of seizure morphology (Gotman et al., 1997). The separate algorithms were developed to detect rhythmic discharges, multiple spikes and very slow rhythmic discharges. The Gotman algorithm inspects the fast Fourier transform (FFT) of an EEG window. Background EEG is pseudo-random, so the FFT power is spread across the spectrum. In contrast to this, seizure EEG consists of a large dominant peak due to the rhythmic morphology. The frequency and bandwidth of the dominant peak frequency of a 10-second window of EEG are compared to a window from 60 seconds in the past. A power ratio between these two samples is compared to a threshold to determine if this is a seizure epoch. This approach mimics the clinical practice of scrolling back in time when analysing EEG to check if the event under investigation differs from the EEG background or mutates over time.

Other feature-based approaches focused on the extraction of features from chaos theory and time-frequency based autocorrelation analysis (Faul et al., 2005; Tapani et al., 2019), based on the understanding that a seizure would be more ordered than surrounding background EEG. These complex features are often well correlated with seizures in neonatal EEG, but they can increase the time required to complete feature extraction.

Model based approaches to detect neonatal seizures have also been developed (Celka & Colditz, 2002; Roessgen et al., 1998). In these approaches a mathematical model of neonatal

EEG is developed. The properties of unseen instances of neonatal EEG are compared with the developed model and seizures are detected based on threshold.

3.2.2 Machine learning based algorithms: the data-driven approach

More recent years have seen the incorporation of machine learning algorithms into neonatal seizure detection research. These algorithms typically have a feature extraction routine followed by a machine learning based classifier. Unlike earlier approaches which utilise empirical human-tuned thresholds to classify EEG (Celka & Colditz, 2002; Gotman et al., 1997; Liu et al., 1992) the rules and thresholds in these algorithms are data-driven. Machine learning algorithms are trained using input examples; they do not need to be explicitly programmed. The learning process typically results from iterative improvement with respect to some cost function over the set of training examples.

Supervised machine learning requires ground truth labels for each training example. The dataset, D, consists of (x_i, y_i) pairs, where $x \in X$ are the input features and $y \in Y$ are the associated ground truth labels. A function, $f(X)$, which maps the input data to the associated labels is learned. This function is then utilised in the inference stage to predict the label of an unseen input datum, $\hat{y}_i = f(x_i)$. In unsupervised machine learning algorithms, underlying patterns and relationships in data are learned without associated training labels. Typically, supervised learning is employed to develop neonatal seizure detection algorithms, this requires labelled training data (Aarabi et al., 2006; Ansari et al., 2016; Tapani et al., 2019; Temko et al., 2011a; Thomas et al., 2013).

In data-driven machine learning neonatal seizure detection algorithms, EEG signals are represented by a set of features which describes the signal characteristics. These features usually represent a segment or window of the temporal EEG; by shifting the window a time-varying representation of the signal is produced. Features are chosen to be informative and to represent a compressed form of the input data which is relevant to the desired task. This has the effect of reducing the required computational complexity of the machine learning model being utilised and highlighting the important characteristics in the data.

Since typically a large set of features is extracted, there is much less emphasis on having a single “magic” feature which discriminates entirely between seizure and non-seizure windows. The features are optimised for their ability to distinguish between seizure and non-seizure patterns and the feature set represents the important characteristics of the data which are required to distinguish between classes.

For some types of machine learning classifier, the inclusion of many features, including irrelevant and redundant features, can inhibit the classification performance and therefore careful feature selection is required (Aarabi et al., 2006). Reducing the number of features in a classification algorithm allows for clearer data visualisation, reduces algorithm memory and computational requirements, and reduces the complexity of the classification model (Guyon & Elisseeff, 2003). Sometimes individual features are selected with a performance metric in mind; for example, a subset of features which maximise accuracy or AUC could be selected from a large feature set. The performance of an individual feature is typically measured by applying a binary threshold to feature values and comparing with the true class labels, or on the separation between seizure and non-seizure EEG windows when they are ordered based on feature values (Greene et al., 2008). Sets of features can also be selected based on the performance of a classifier trained using different subsets of the overall feature set. Feature transformation techniques, such as principal component analysis and linear discriminate analysis, can also be utilised to reduce the dimensionality of features and decorrelate the feature space, or to transform the feature space to optimise the separation of classes (Thomas et al., 2013). Feature selection and dimensionality reduction simplify the required classification model which can then lead to shorter training times and can reduce classifier overfitting.

Machine learning models can be classed as either generative or discriminative. A generative classifier seeks to model how the data was generated, by learning the joint probability distribution $P(X, Y)$. A generative model will then utilise some means to understand which class is more likely given an unseen input example. In contrast to this, a discriminative classifier will directly learn a conditional probability of the label Y given the observation X, $P(Y|X)$ (Ng & Jordan, 2002). A discriminative classifier makes no assumptions about the underlying distribution of the data.

The k -nearest-neighbours classification algorithm (KNN) is an example of a discriminative non-parametric supervised machine learning model. All training examples are mapped in the feature space, during the inference stage the class label of the k nearest neighbours to a new unseen example are utilised to predict the class. This prediction can then be made in a variety of ways, including majority voting or a weighted average of the neighbours based on distance, d, of each of the k nearest neighbours. Nearest neighbour based classifiers have been proposed for seizure detection in adults (Qu and Gotman, 1997).

Gaussian Mixture Models (GMM) are parametric generative machine learning models. The algorithm assumes that probability density function of data can be modelled by a mixture of Gaussian distributions in the feature space. Each Gaussian distribution, Equation 3.1, is represented by its mean μ, which defines its centre and covariance $\sum$, which defines its width and each one is weighted in the mixture by some value w.

$$g(x|\mu_j, \Sigma_j) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_j|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)\right) \quad (3.1)$$

Typically, each class is represented by a separate model. The Gaussian parameters and the weights are found through algorithm training. Expectation maximisation is a commonly employed algorithm, which iteratively finds new sets of parameters to represent the datapoints (Dempster et al., 1977). During the testing stage the likelihood estimate is generated for each class based on a single test example. The Bayesian formula is utilised to yield the posterior probability of each class for the test example. GMMs have been utilised for neonatal seizure detection (Thomas et al., 2010).

Logistic regression, support vector machines (SVM) and neural networks (NN) are examples of discriminative machine learning classifiers. These algorithms directly learn how to classify an input signal without attempting to model the distributions of the data. Two examples of discriminative classifiers, SVMs and NNs, will be explored in more detail later in this chapter.

3.3 Neonatal seizure detection baseline algorithms

Two data driven neonatal seizure detection algorithms will be utilised as baselines in this thesis. These algorithms have both been developed to detect seizures in term infants. Both algorithms rely on SVM classifiers to detect seizure events. This section will briefly introduce the theory behind SVM classifiers and then present the Cork SVM algorithm and the Helsinki SVM

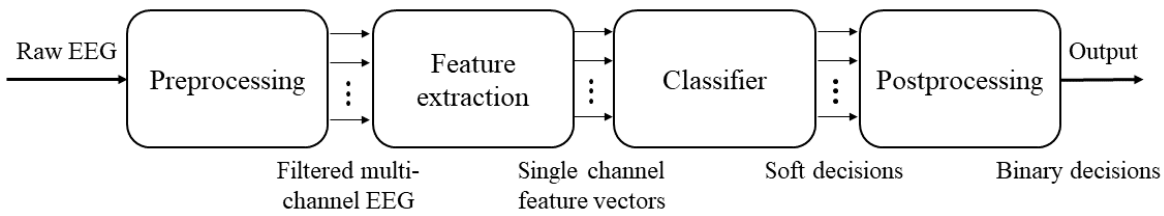


Figure 3.1 General architecture for the SVM based neonatal seizure detection algorithms. Both the Cork SVM and the Helsinki SVM neonatal seizure detection algorithms follow this structure and utilise SVM classifiers.

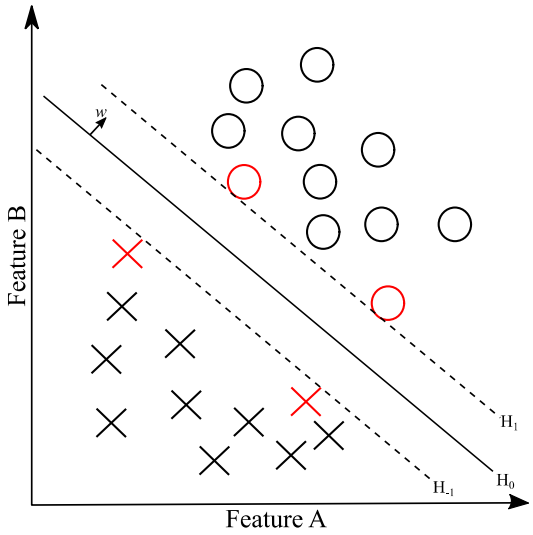


Figure 3.2 SVM decision boundary with support vectors marked in red for linearly separable classes. The optimal support hyperplane is H_0, the supporting hyperplanes are H_1 and H_{-1}.

algorithm. Figure 3.1 shows a block diagram which describes the main stages of both the Cork SVM and Helsinki SVM algorithms.

3.3.1 SVM theory

For a period from 2011 to 2019 a combination of medically informed feature engineering and the SVM classification algorithm was a popular and highly effective formula for developing neonatal seizure detection algorithms (Ahmed et al., 2017; Tapani et al., 2019; Temko et al., 2011a). This framework required that neonatal EEG be represented as a set of informative features. The nature of the SVM algorithm means that it is not very sensitive to redundant features (Weston et al., 2000). This implies that relatively large feature sets could be utilised, and minimal information about the neonatal EEG is lost when EEG is translated to a large set of representative features.

The SVM is a popular discriminative supervised learning model which can be used for both classification and regression tasks. The power of the SVM model comes from the assumption that a transformation into a higher dimensional space can simplify a complex problem into one that can be solved using a linear separation function. SVMs also benefit from the fact that only a subset of the training samples, the support vectors, are required during the inference stage. These support vectors are the training examples that are closest to the decision boundary, the assumption is that these provide the most important information for unseen sample classification.

An SVM can be initially represented in the context of examples which are linearly separable. A representation of a binary problem with linearly separable data is shown in Figure 3.2, the aim is select a hyperplane which separates the classes. The classification of each input vector, $x_i \in \mathbb{R}^n$, can be defined as,

$$f(x_i) = w \cdot x_i + b. \tag{3.2}$$

A corresponding set of labels are available for each input vector, where $y_i \in \{1, -1\}$. The hyperplane H_0 is the region of vectors which satisfy $f(x) = 0$; where, $w \in \mathbb{R}^n$ and $\cdot$ denotes the dot product of two vectors. This linear classification function is parameterised by w and b: these parameters are selected to maximise the distance (margin) between this dividing hyperplane and the data in each class. The decision boundary, the hyperplane H_0 in Figure 3.2, separates the data into two classes and it maximises the distance to both classes.

To be correctly classified the vectors must satisfy

$$y_i(w \cdot x_i + b) \geq 1 \ \ \forall i. \tag{3.3}$$

Two more hyperplanes, H_{+1} and H_{-1}, define the margin between the data classes. These hyperplanes are parallel to the decision boundary, they intersect the data points in each class which are closest to the decision boundary. These hyperplanes are defined by $H_{+1}: f(x) = 1$ and $H_{-1}: f(x) = -1$. The distance separating the supporting hyperplanes, known as the margin, is $d = \frac{2}{\|w\|}$, the classifier is optimised by maximising this distance; this implies a minimisation of $\|w\|$. Maximising the distance, d, ensures that the classifier has better generalisation ability when classifying future, unseen data. The problem can be reframed as an optimization task where $\frac{1}{2}\|w\|^2$ is minimised subject to the constraints in Equation 3.3. This optimisation problem has a unique solution.

Often, real-world datasets are affected by outliers and sometimes noisy or incorrect samples are included, or the classes in a dataset may not be completely separable. The presence of these samples may reduce the margin of separation and reduce the classifier's generalisation ability. To overcome this the decision boundary can be softened by including a slack variable, ξ_i. If this slack variable is too large for each training example, then the optimisation will result in a trivial solution, so a penalty on the size of ξ_i is included in the objective function. The problem now becomes:

minimise $$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{L}\xi_i, \quad (3.4)$$

subject to $$y_i(w \cdot x_i + b) \geq 1 - \xi_i \ \forall i. \quad (3.5)$$

This quadratic programming optimisation problem can be solved through Lagrange multipliers to give a unique optimal solution. The training examples which do not lie on the decision boundary will have optimal value of 0, the support vectors are the training examples which lie on the supporting hyperplanes. For a set of M support vectors, the SVM classifier is:

$$f(x_i) = \sum_{j=1}^{M}\alpha_j y_j x_j \cdot x_i + b \quad (3.6)$$

where $$\sum_{j=1}^{M}\alpha_j y_j = 0 \text{ and } \alpha_j > 0. \quad (3.7)$$

If the data is not linearly separable then a mapping function can be used to transform the data into a higher dimensional representation. In this instance the mapped version of the data, $K(x_i, x_j)$, replaces the original data; this mapping is known as a kernel function. The inclusion of the kernel function into the SVM algorithm changes the classifier decision to:

$$y(x_i) = \text{sgn}\left(\sum_{j=1}^{M}\alpha_j y_j K(x_i, x_j) + b\right) \quad (3.8)$$

where the sgn() function determines the classification decision based on the sign.

The SVM algorithm can represent complex functions while remaining resilient to overfitting. A key insight of the SVM optimisation process is that the algorithm expected generalisation loss is minimised (maximal margin separator), rather than minimising the empirical loss on the training data (Boser et al., 1992). This also gives the benefit of a single global minimum due to the convex loss function.

3.3.2 Cork SVM algorithm

The Algorithm for Neonatal Seizure Recognition (ANSeR) is an SVM based seizure detection algorithm, this will be referred to as the Cork SVM in the remainder of this thesis. The algorithm has been validated in a multi-stage international clinical trial (Mathieson et al., 2016; Temko et al., 2011a). It was successfully validated in eight NICU's across Europe, no other neonatal seizure detection algorithm has been as robustly tested in this kind of clinical environment. The algorithm has been licenced to an international medical device company.

The development of the Cork SVM algorithm took many years and harnesses the work of many signal processing, engineering, machine learning, and medical experts. There are, however, some potential limitations in the Cork SVM algorithm. One limitation is the independent optimisation of the feature extraction stage and the classifier. The requirement for a feature set may also represent a limitation as there may be some information in the EEG signal which is not captured in the feature set. A second limitation of this algorithm is the requirement for per-channel seizure annotations, there is a large clinical cost associated with the generation of channel specific seizure annotations.

The Cork SVM algorithm was widely seen as a robust neonatal seizure detection algorithm. It achieved a leave one patient out (LOO) testing performance of 96.6% AUC on a dataset of long EEG recordings (Temko et al., 2013). This was considered to be the state-of-the-art in neonatal seizure detection for many years. One key takeaway from this work is that when many relatively simple EEG signal features are combined with a powerful and robust classifier, such as the SVM, a reliable detector can be trained.

a) Preprocessing

In the Cork SVM algorithm the raw EEG signal is first preprocessed. EEG is typically sampled at frequency rates from 256Hz up to 1024Hz. These relatively high frequency recordings mean that a large amount of the recorded frequency range is not useful for the seizure detection task; the majority of seizure power is in the δ and θ sub-bands (Kitayama et al., 2003). In the Cork SVM algorithm, the EEG signal is digitally band pass filtered between 0.5Hz and 12.8Hz. This uses a low pass filter constructed from a 65-coefficient finite impulse filter with a Hanning window. Low pass filtering ensures that the 50Hz electrical mains noise is dramatically reduced and of course also acts as an anti-aliasing filter. The high pass filter is a first order infinite impulse response filter, it removes the slow drift of the EEG. Filtering removes low frequency artifacts and irrelevant high frequency information, while retaining the dominant seizure frequencies. These filtered EEG signals are then downsampled to 32Hz to simplify the representation while still maintaining the vital information required to detect seizures.

An epoch length of 8-seconds with a 4-second overlap is utilised in the Cork SVM algorithm. This choice is based on the knowledge that seizures have a minimum duration of 10-seconds, according to the definition of a neonatal seizure event (Clancy & Legido, 1987). The minimum seizure epoch length possible would be 10-seconds, to ensure that the window of interest will contain only seizure activity. Also, seizure activity is mostly concentrated in the range of 0.5-

Table 3.1 The Cork SVM algorithm feature list is made up of 55 features which have been carefully engineered and optimized to give best performance on the neonatal seizure detection task.

	Cork SVM Features
Frequency Domain	• Total power (0-12Hz) • Peak frequency of Spectrum • Spectral edge frequency (80%, 90%, 95%) • Power in 2Hz wide sub-bands • Normalized power in sub-bands • Wavelet energy
Time Domain	• Curve length • Number of maxima and minima • Root mean squared amplitude • Hjorth parameters (activity, mobility, complexity) • Zero crossings (raw epoch, Δ, ΔΔ) • Autoregressive modelling error (model order 1-9) • Skewness • Kurtosis • Non-linear energy • Variance (Δ, ΔΔ)
Information Theory	• Shannon entropy • Singular value decomposition entropy • Fisher information • Spectral entropy

8Hz (δ and θ bands); in order to capture a full period of seizure activity at 0.5Hz then at least a 2-second window would be required. However, the definition of seizure involves the search for repetitive patterns, so the ideal epoch window would be a multiple of this minimum window size. Due to the tendency of seizure patterns to morph over time, relatively small windows of EEG should be used to search for repetition, this also limits the maximum size of the chosen EEG epoch. Epochs of 8-seconds sampled at 32Hz also have the benefit of producing 2^8 samples in a window, this makes any FFT more efficient.

b) Feature extraction

The Cork SVM algorithm uses a large feature set to represent each 8-second window of EEG. The features are taken from the time, frequency, and information theory domains. This feature set was optimised for the neonatal seizure detection task and many years of engineering, machine learning and medical expertise went into selecting the 55 features which demonstrated potential for the neonatal seizure detection task. Table 3.1 shows the extracted feature set in more detail. Features are extracted from the preprocessed EEG signal and are normalised by

subtracting the mean and dividing by the standard deviation to assure commensurability of various features.

The set of frequency domain features are used to describe the changes in the frequency spectrum of EEG windows. Some of the frequency domain features are calculated for the complete signal. Others are utilised to describe smaller 2Hz wide frequency sub-bands; those sub-bands are: 0-2Hz, 1-3Hz, 2-4Hz, 3-5Hz, 4-6Hz, 5-7Hz, 6-8Hz, 7-9Hz, 8-10Hz, 9-11Hz and 10-12Hz. Most of these frequency domain features rely on the computation of an FFT for each EEG window, this assumes that the EEG signal is stationary.

The set of time domain features describe the morphology of the signal. Some of these features describe statistical characteristics of the raw signal, others are based on the first (Δ) and second ($\Delta\Delta$) derivatives of the signal. The set of information theory features measure the order and predictability of each EEG window. Seizures are known to be more rhythmic than background EEG; entropy measures can be utilised to measure the uncertainty of a signal.

c) Classifier

The Cork SVM algorithm utilises an SVM classifier, this classifier is applied separately to each channel of neonatal EEG. Per-channel (strong) seizure annotations are required for training because neonatal seizure labels must be localized to a single EEG channel to train the single channel classifier. A Gaussian kernel was used in the SVM. Five-fold cross-validation was used on the training data to find the optimal set of parameters for the Gaussian kernel and for the generalization parameter. On average around 19% of the per-channel annotated training data became support vectors. The most chosen Gaussian kernel and generalization parameter selected through the cross-validation routine were $\gamma = 0.05$ and $C = 20$.

Once the best set of hyper-parameters was chosen, the model was trained using all the training data. A leave-one-out cross-validation testing procedure was used to test the classifier's performance as a patient independent seizure detection system. The classifier was then applied separately to each EEG channel of the test patient. The test outputs from each of the 8 channels are converted to probabilistic values using a sigmoid function based on the probabilistic priors in the training dataset (Platt, 1999).

d) Postprocessing

The final stage in a neonatal seizure detection algorithm is the postprocessing routine, where the raw probabilistic outputs are combined and smoothed over time. The outputs of the SVM

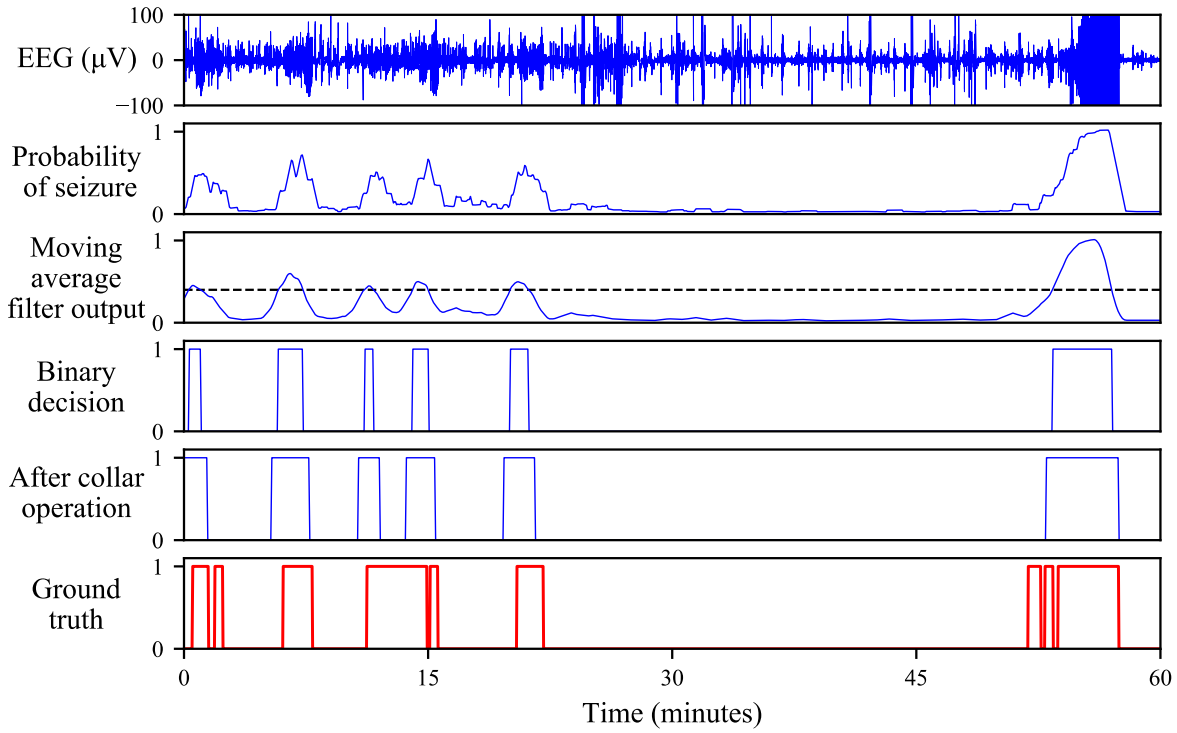


Figure 3.3 Example showing the effects of different stages of the Cork SVM algorithm postprocessing on a real EEG example. The EEG signal and the algorithm decision output at each stage of postprocessing are shown in blue, the ground truth seizure labels are shown in red.

classifier are typically converted to probabilities using Platt scaling (Platt, 1999); zero represents a non-seizure window and one represents a seizure window. The postprocessing algorithm is informed by clinical knowledge about the characteristics and the presentation of neonatal seizures. A selection of postprocessing steps are presented in Figure 3.3.

The first step in the postprocessing routine is a moving average filter, which has a smoothing effect on the probabilistic output. A 60-second moving average filter is applied to the probabilistic output, this diminishes values which diverge from the moving mean. Clinical neurophysiologists will often review the EEG trace directly before and after a suspected seizure event, this is mimicked though applying a moving average filter which is influenced by the surrounding seizure probabilities. The moving average filter is a non-causal filter, it requires 30 seconds of information after the epoch being processed to produce a decision. For the application of neonatal seizure detection this delay is deemed to be acceptable.

The threshold operation transforms the continuous probabilistic output into binary decisions. The binary decisions of all the available EEG channels are then combined into a single multi-channel output. Clinical neurophysiologists are interested in the presence of seizure events; they want to know if there is a seizure event happening in any of the channels. The binary OR

operator allows engineers to search for a seizure in any one channel and report this as a multichannel output. Applying the OR operator across all channels produces a single binary decision vector.

A collar is applied to the multichannel binary decision vector, this extends all detected seizure events forward and backward in time. In this work, seizure decisions are extended by 30 seconds in each direction (half of the length of the moving average filter), this is to account for the delays introduced by the moving average filter.

The final stage in the postprocessing routine is a retrospective artefact removal operation (Temko et al., 2013). The respiration artefact is a particularly seizure-like signal which can be mistaken for seizures by automated seizure detection algorithms. An adaptive threshold routine has been developed to overcome this challenge. The operation is performed for each suspected seizure event that passes the threshold. Over a 600 second interval the average probability of non-seizure epochs is calculated. This is subtracted from the probability for each epoch in the suspected seizure event; these epochs are only considered to be seizure epochs if this adapted probability is greater than the threshold.

3.3.3 Helsinki SVM algorithm

The Helsinki algorithm is an SVM based neonatal seizure detection algorithm which uses a set of 21 features (Tapani et al., 2019). The development of the SVM Helsinki algorithm saw the development of two bespoke features designed specifically for the neonatal seizure detection task. The Helsinki SVM algorithm is an example of the power of feature engineering; the authors show that the developed features are well correlated with neonatal seizures meaning that the individual features have relatively high AUC scores when tested against seizure and non-seizure window labels.

a) Preprocessing

The algorithm is applied to each channel separately, this implies that channel-specific (strong) annotations are required for algorithm training. Each EEG channel is high pass filtered with a cut-off frequency of either 0.5Hz or 1Hz; see Table 3.2 for more information on how the selective high pass filter cut-off frequencies are utilised. A 50Hz notch filter is applied to remove electrical noise and a 30Hz low pass filter is applied to remove high frequency noise. The EEG is down sampled from 256Hz to 64Hz and segmented into 32s windows with a 4s shift between successive windows.

b) Feature extraction

Table 3.2 The Helsinki algorithm feature set is made up of 21 features. This table shows the high pass filter values in preprocessing for the features. All features with an asterisk (*) are optimised for preprocessed EEG with high pass filter with cut-off frequency of 0.5Hz, all other features use a high pass filter with a cut-off frequency of 1Hz.

	Helsinki SVM Features
Frequency Domain	• Log of spectral power (0.5 – 30Hz) • Median correlation of TFC • *Total harmonic distortion 1 (power in the first three harmonics divided by the sum of the power spectral density) • *Total harmonic distortion 2 (power in the fundamental frequency divided by the sum of the power spectral density) • *Total harmonic distortion 3 (log of the power in the fundamental frequency) • *Relative power in the delta band (0.5 – 4Hz) • *Relative power in the theta band (4 – 8Hz) • *Relative power in the alpha band (8 – 12Hz) • *Relative power in the beta band (12 – 30Hz) • *Total harmonic distortion of the time frequency distribution
Time Domain	• Skewness of SNLEO • Regularity of 2s sub-windows of SNLEO • Spike number • Spike median width • Spike median gap • Mean SC • Std SC • Mean of SNLEO • Std of SNLEO • Mean spike peaks over background • Amplitude envelope

The Helsinki SVM algorithm uses a set of 21 features to represent each 32-second window of EEG. Two features were explicitly developed for this algorithm; both novel features are focused on detecting correlation in the non-stationary EEG signal. The full feature set is shown in Table 3.2.

The first bespoke feature measures spike correlation (SC); a maximum cross-correlation with respect to time lag between adaptively extracted segments of EEG. To calculate the SC of an EEG window the signal is adaptively segmented into spike events. The distinction between spike events is found based on the smoothed Non-Linear Energy Operator (NLEO) of the signal. The correlation between pairs of spikes in each window is calculated using a type of

normalised autocorrelation function. The average SC and standard deviation of SC within an EEG window are utilised as features in the machine learning algorithm.

A second bespoke feature is the time frequency correlation (TFC), this is the maximum cross-correlation between scale-shifted time slices of a time-frequency distribution. A time frequency distribution characterises signal energy in the frequency domain across a range of successive time-steps. Time frequency distributions can be utilised to demonstrate how the frequency components of a signal shift over time. To calculate the TFC feature the maximum correlation of scale shifted time slices of the time frequency distribution are taken for an EEG window.

c) Classifier

An SVM algorithm is utilised to the classify feature vectors for each 32-second window of EEG. Similar to the Cork SVM algorithm, one SVM algorithm is applied to each individual channel of EEG. The SVM was trained using 5-fold cross-validation and Bayesian optimisation within the cross-validation loop (Optimize a Cross-Validated SVM Classifier Using Bayesopt - MATLAB & Simulink, 2021). Again, the features are normalised by subtracting the mean and dividing by the standard deviation. The Helsinki SVM utilises a Gaussian kernel.

d) Postprocessing

The output of the SVM for each EEG channel is a continuous value which indicates the distance from the learned SVM hyperplane. A three-sample moving average filter is applied to the SVM output trace. A maximum operator across the channels transforms the multi-channel outputs into a single continuous temporal value. This is further processed with a filter which calculates the median value of every three samples. A threshold is applied to this value to generate a vector of binary seizure or non-seizure decisions.

3.4 Performance qualification, metrics, and model generalisation

Performance metrics allow for the comparison of seizure detection classifiers and give an indication of how the developed algorithms would perform in a realistic NICU environment (Temko et al., 2011b). Epoch-based metrics are the most reported in literature, but event-based metrics give an important representation of classifier performance from a clinical perspective. Performance metrics also give an indication of the optimal threshold to select for a classifier to give a desired level of sensitivity or specificity. All classifiers developed in this work will return a probabilistic output and a threshold must be applied to this output to classify epochs as either seizure or non-seizure.

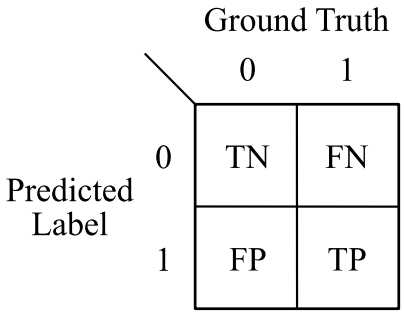


Figure 3.4 Confusion matrix which shows the performance of an algorithm by classifying the errors and the correct predictions.

To generate epoch-based performance metrics the full binary decision vector, after a threshold has been applied to the probabilistic output, is compared with the corresponding ground truth vector. Each epoch is labelled as true positive (TP), false positive (FP), true negative (TN), false negative (FN). A confusion matrix showing the relationship between ground truth labels and predicted labels is presented in Figure 3.4. The classifier sensitivity and specificity are calculated from the sum of epochs for each label.

Classifier sensitivity is the percentage of seizure epochs correctly labelled:

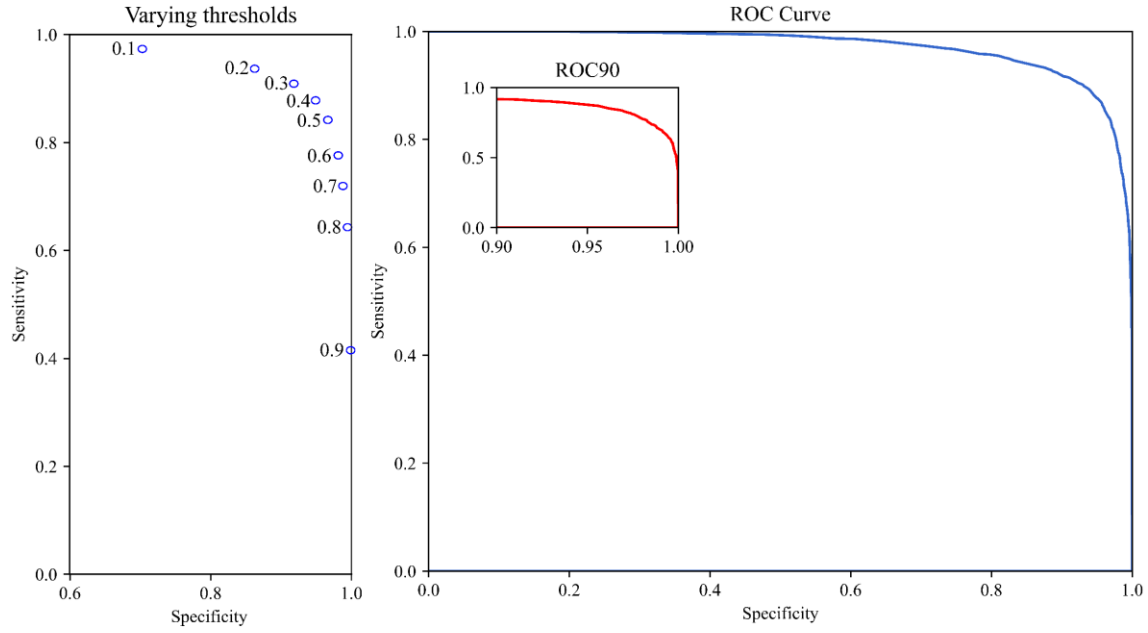


Figure 3.5 On the left the varying sensitivity and specificity for different classification thresholds is shown. The right figure shows an example AUC curve and AUC90 curve. The curve is generated by plotting sensitivity against specificity for a range of threshold values. The area under the ROC curve gives an indication of classifier performance. The AUC90 metric indicates the performance of an algorithm for thresholds which give very high levels of specificity. The dashed line represents a test where there is no separation between classes and the resultant AUC is 0.5.

$$Sensitivity = \frac{\sum TP}{\sum TP + \sum FN} \tag{3.9}$$

Classifier specificity is the percentage of non-seizure epochs correctly labelled:

$$Specificity = \frac{\sum TN}{\sum TN + \sum FP} \tag{3.10}$$

The receiver operation characteristic (ROC) graph is a plot of the sensitivity and specificity pairs resulting from varying the decision threshold over the range of classifier outputs observed. In most cases, for high threshold values, close to the maximum observed output, the sensitivity will be close to one, but the specificity will be close to zero. The inverse would be true for low threshold values. The area under the ROC (AUC) measures the separation of the classes. The closer this area is to one, the better the separation between the binary classes, the better the classifier's performance. An example ROC curve is shown in Figure 3.5.

The AUC provides a single statistical value which can be used to compare the epoch-based performance of classifiers. An AUC value of 0.5 implies that there are no apparent differences between the classes, the performance is no better than a random guess. The AUC quantifies the probability that a random epoch with a seizure label will be assigned a higher probability than a randomly selected non-seizure epoch (if a label of one indicates a seizure). Metrics such as accuracy are reliant on the distribution of each class in a dataset and on the threshold selected. These threshold-based metrics represent a single point across the ROC, whereas the AUC represents the overall separation between the class labels.

For a seizure detection system to be clinically usable, the specificity needs to be high to result in an acceptable number of false detections per hour. The AUC90 metric is the area under the curve for specificities over 90%. An example ROC90 curve is shown in Figure 3.5. This is a more clinically relevant performance measure, as it looks at the amount of seizure epochs detected for low levels of false seizure detections; low specificity would result in unwanted false alarms in the NCIU.

In this work, a seizure event is recorded as detected if there is any overlap between a positively labelled seizure event and the ground truth event label. A correct detection is any overlap of a predicted positive seizure label with a seizure event in the ground truth. Any detection that does not overlap with a true seizure event is counted as a falsely detected (FD) event. The FD/h metric indicates the number of times that a clinician will have to check the results of a seizure detection algorithm in vain; this is an unwanted consequence of employing a seizure detection

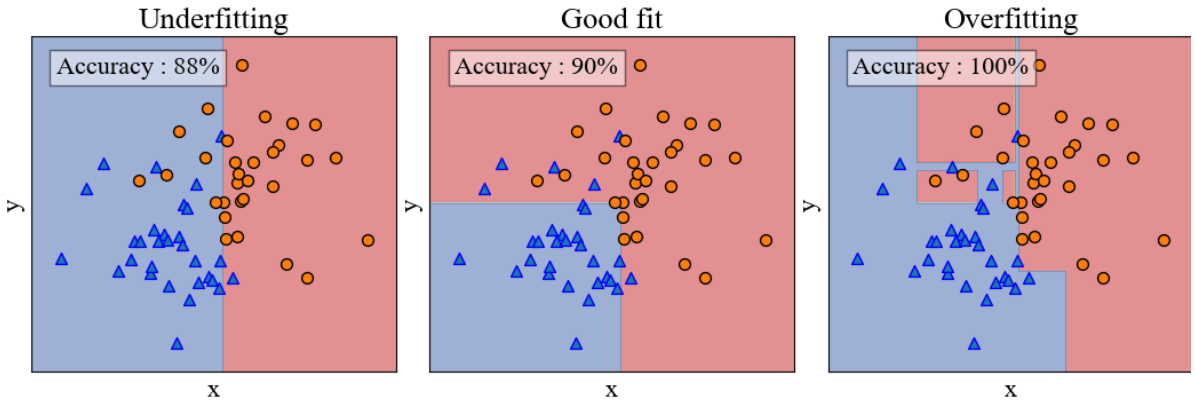


Figure 3.6 Visual representation of the three different levels of model complexity fit to the same input data. The model that is underfitting is not complex enough to represent the data. The model that is overfitting is too complex and fits to the noise in the dataset.

algorithm in the NICU. This metric is clinically relevant because clinical staff will only trust an algorithm if the FD/h is at a tolerably low level (Temko et al., 2011b).

The aim of supervised machine learning is to train a function $f(x)$ which will predict output value $\hat{y}$ given input x. This mapping is learned using a training dataset and the aim is for the algorithm to exhibit good performance when tested on an unseen testing dataset, the measure of good performance is dependent on the chosen performance metric. The ability for a trained model to generalise is important because a training dataset is only a sample of all possible data; datasets contain noise and are incomplete. Two important sources of error that can inhibit the generalisation ability of a model are underfitting and overfitting. Examples of underfitting and overfitting are presented in Figure 3.6.

Overfitting occurs when the learned function $f(x)$ is mapped too closely to the training data. A model with too much learning capacity tends to learn the detail and noise in the training dataset, this will negatively affect the model's performance on unseen data. Overfitting can be addressed by limiting the model's learning capacity or stopping early in the training phase if it is an iterative training process.

Underfitting occurs when the learned function $f(x)$ cannot sufficiently model the training data and in turn cannot model the unseen testing data. To overcome underfitting a more complex model should be employed to more accurately predict the data.

3.5 Introduction to deep learning

In recent years, the use of classical machine learning techniques in areas such as speech, language and image processing has declined, while deep learning-based algorithms have seen

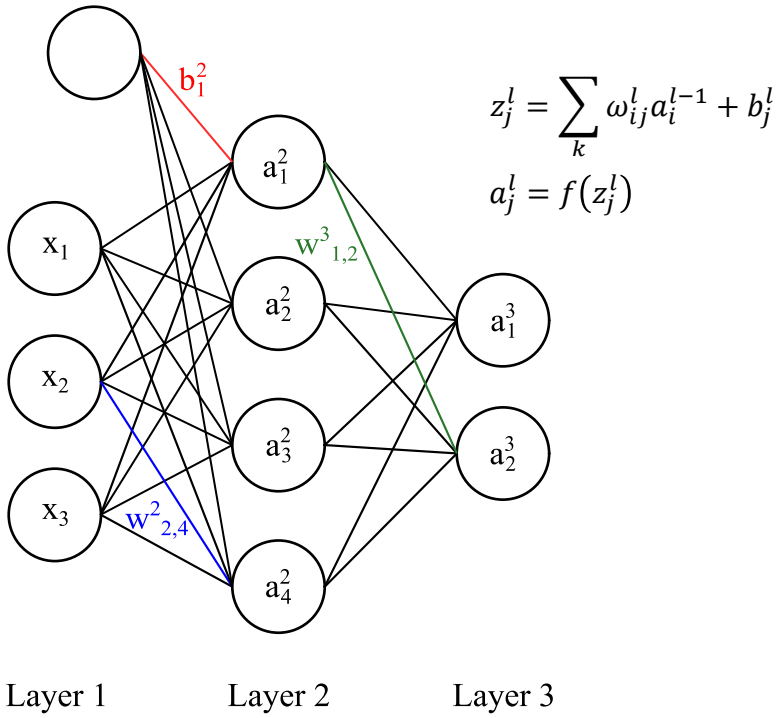


Figure 3.7 An example MLP architecture. All layers are densely connected i.e. all neurons in a layer are connected to all neurons in the next layer, but there are no connections between neurons within a layer. The nodes in each layer take a weighted sum of their inputs and apply a non-linear function to this value, the result becomes the input to the subsequent layer. In this example, Layer 1 is the input layer, Layer 2 is the hidden layer and Layer 3 the is output layer.

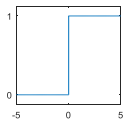
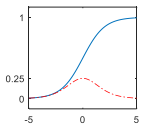
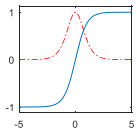
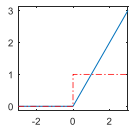
a surge in popularity. Deep learning algorithms use multiple layers to hierarchically extract increasingly complex internal representations of network inputs; these learned representations are used to predict an output value. Some deep learning algorithms can learn task specific internal representations from untransformed input data, this eliminates the need for feature engineering. Removing the feature engineering stage limits the influence of certain human induced biases in algorithm development.

3.5.1 The multilayer perceptron

The multi-layer perceptron (MLP) is made of perceptron neurons organised into layers, which are connected in a feedforward, fully connected fashion, with no intra-layer connections. The depth of an MLP is measured by the number of hidden layers; the input, hidden and output layers can be seen in Table 3.3.

The addition of hidden layers means that the MLP can model data that is not linearly separable, distinguishing it from the perceptron, which consists of feed forward, fully connected input and output layers with no non-linear operations. The non-linear function is an important attribute of the MLP network, without this non-linear function the weighted multiplications in each layer

Table 3.3 Common activation functions used in deep learning networks. The functions and their derivatives are plotted in the final column.

Non-linearity	Function	Derivative	function / derivative
Binary step	$f(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0, & x \neq 0 \\ ?, & x = 0 \end{cases}$	
Sigmoid	$f(x) = \frac{1}{1+e^{-x}}$	$f'(x) = f(x)\big(1 - f(x)\big)$	
Tanh	$f(x) = \frac{2}{1+e^{-2x}} - 1$	$f'(x) = 1 - f(x)^2$	
ReLU	$f(x) = max(0, x)$	$f'(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$	

of the network could be reduced to a single operation i.e. a model with no hidden layers. Therefore, the non-linearity is the source of network depth. The most common non-linear functions (also known as activation functions) used in deep learning are the sigmoid, the hyperbolic tangent and the rectified linear unit (ReLU). Four common non-linear activation functions are shown in Table 3.3 alongside their derivatives.

The value for each layer in the network is calculated using Equations 3.11 and 3.12, where a_j^l is the value of the j^{th} neuron in the l^{th} layer. The weight, w_{ij}^l, connects the i^{th} neuron in the $(l-1)^{th}$ layer with the j^{th} neuron of the l^{th} layer. For the j^{th} neuron in the l^{th} layer, z_j^l is the weighted summation of outputs from all the neurons in the previous $(l-1)^{th}$ layer. A non-linear function is applied to z_j^l to calculate the neuron's output, a_j^l. These equations and their relationship to the neurons in feedforward layers are also shown in Figure 3.7.

$$z_j^l = \sum_i w_{ij}^l a_i^l + b_j^l \tag{3.11}$$

$$a_j^l = f(z_j^l) \tag{3.12}$$

The loss function measures the performance of a classifier by comparing the classifier's predictions, $\hat{y}$, with a known ground truth, y, for each training example. $\hat{y}$ is the vector of all values in the final layer of the MLP given by $[a_1^L, a_2^L, a_3^L \ldots . a_j^L]^{\mathrm{T}}$, where L is the number of

layers in the network. A loss function must be non-negative. The output of a loss function should tend towards zero as the network predictions become more accurate. The quadratic loss (mean squared error) is a commonly employed loss function, which is given by

$$E(w,b) = \frac{1}{2n}\sum_{x}\|y(x) - \hat{y}\|^2. \tag{3.13}$$

The cost is a function of the network parameters (weights and biases or w and b) and is summed over all N_p pairs in a training set, each training pair being (x_p, y_p). The cross-entropy loss function is very commonly used in classification problems. The target labels should be probabilistic; targets should be a number between zero and one. The binary cross-entropy loss is defined by

$$E(w,b) = -\frac{1}{N_p}\sum_{p}\left[y_p \ln \hat{y}_p + \left(1 - y_p\right)\ln\left(1 - \hat{y}_p\right)\right] \tag{3.14}$$

This binary cross-entropy loss is defined for a binary problem, the cross-entropy cost can be adapted slightly to become the categorical cross entropy cost, which can be utilised for multi-class classification tasks. The categorical cross-entropy function is

$$E(w,b) = -\frac{1}{N_c}\frac{1}{N_p}\sum_{p}\sum_{c}\left[y_{c,p}(x) \ln \hat{y}_{c,p}(x)\right], \tag{3.15}$$

where observations are summed over the N_c output classes and the sum of all output classes is always one.

3.5.2 The convolutional neural network

The convolutional neural network (CNN) is a deep learning algorithm which utilises convolutional filters to detect local patterns at varying spatial positions. The neurons take inputs, perform a dot product between the inputs and a weight vector and apply a non-linearity to the result, similar to an MLP network. Again, the output is assessed using a differentiable cost function, in the same way as in an MLP network, but in the CNN the weights are locally shared and are not densely connected.

The key to understanding the CNN design is recognising that they were originally intended for the sole purpose of image processing from raw RGB images (images represented as red, green, and blue channels). In image processing, the importance of spatially local patterns is well understood so CNNs utilise sparse locally connected weights. Image processing also incorporates the knowledge that patterns can appear in any area of an input image; weights are

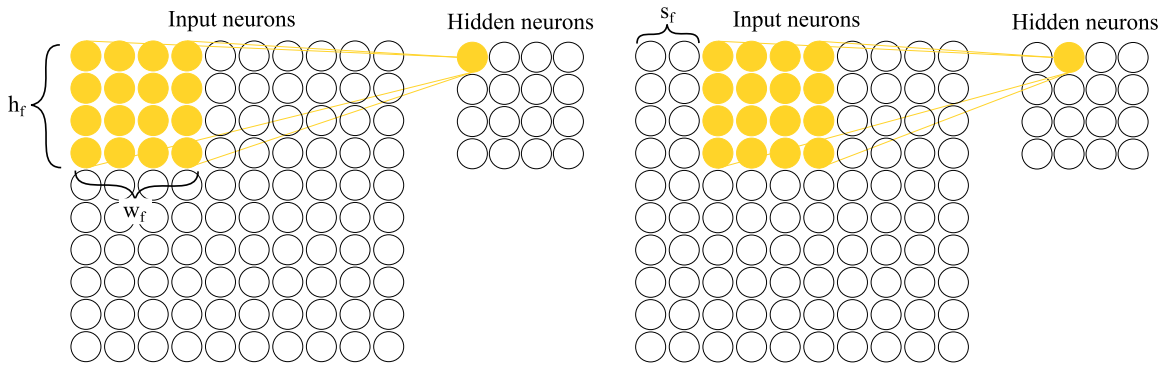


Figure 3.8 Convolving a 4x4 filter across a 10x10 input with filter stride 2 results in a 4x4 feature map. The convolutional weights are locally connected and are shared because the same weights convolve across the input. The effects of filter height ($\boldsymbol{h_f}$), width ($\boldsymbol{w_f}$) and stride ($\boldsymbol{s_f}$) parameters can be seen in the generated feature maps in this example.

designed to be robust to spatial input transformations; this is in contrast to MLPs where the input feature vector is always represented in a fixed order.

Convolutional layer

The **convolutional layer** consists of a set of trainable weights which span a limited area of input width (W) and height (H) but span across all the input channels (M); images are often represented as three (red, green , blue) input channels. The dot product of the weight and the input is calculated in a systematic way across both the W and H dimensions of the input; this is the convolutional operation from which the layer gets its name. The outputs of this convolutional process are stored as a 2-dimensional map, which could be viewed as a transformed representation of the input with spatial relationships maintained. This is called a feature map. Figure 3.8 shows how intermediate feature maps are generated from spatially local weights convolving across an input. A non-linearity is typically applied to the feature map. Each different trainable weight vector in a layer will produce a different feature map. The different feature maps form the channel dimension at the input to the subsequent convolutional layer. Figure 3.9 (a) shows how the dimensionality of convolutional filters is related to the number of input channels and the chosen width and height of the filter.

The convolutional layer hyperparameters control the characteristics of the feature map representation that is generated. The number of convolutional filters controls the number of feature maps. The size of each filter's weight kernel impacts the complexity of the pattern that can be represented in each layer. The stride between successive convolutional kernel operations controls the complexity of the feature map representation by influencing the W and H parameters in the output feature map. Padding also controls the size of the output feature map.

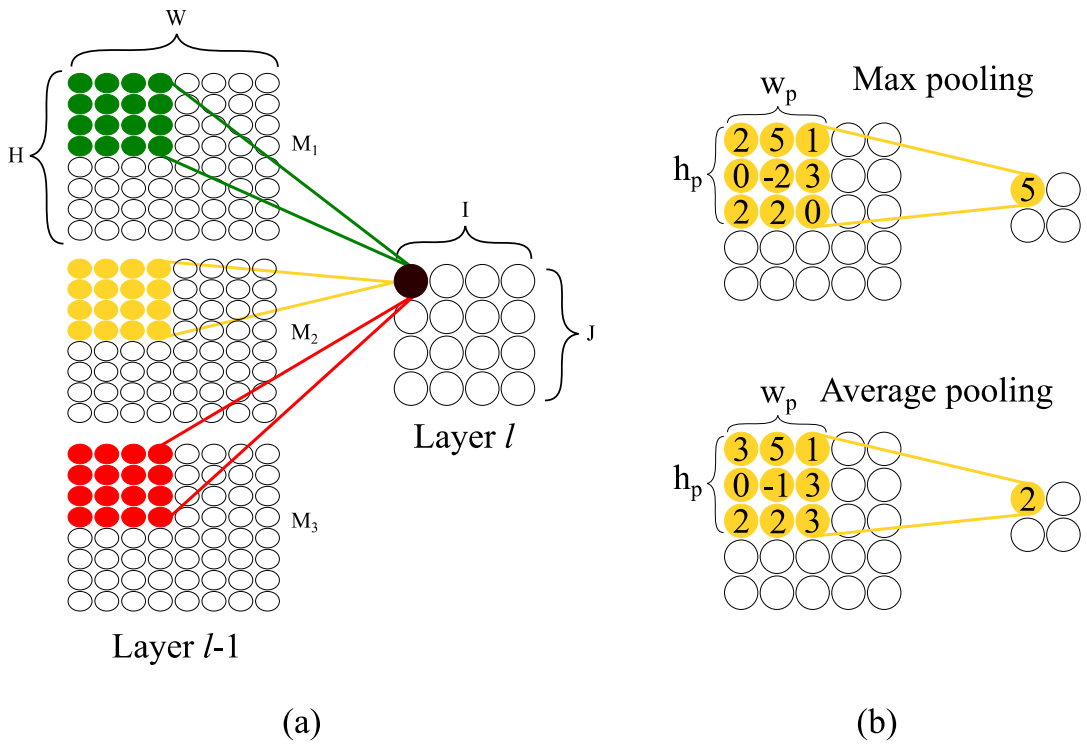


Figure 3.9 (a) A single convolutional filter is applied to all input channels (in the hidden layer these input channels are the feature maps of the previous layer). Convolutional filters typically have three dimensions - number of input channels by width by height. In this example, a convolutional filter was applied to all feature maps in layer l-1 and it generated a value in a feature map in layer l. (b) Pooling acts like a downsampling of the values represented by each individual feature map. The most common pooling operations are max pooling and average pooling.

Pooling layer

The **pooling layer** is a form of downsampling which is applied to feature maps in CNN architectures. Pooling operations are applied to local sub-regions of each feature map; this is illustrated in Figure 3.9 (b). The pooling filter moves across the input in a similar way to the convolutional filter. The size of the pooling filter and the stride between pooling operations determines the extent of the downsampling effect. The most common pooling operations are maximum and average pooling.

Pooling operations downsample the spatial size of feature map representations, which reduces the computational resources needed to implement a CNN. This operation also makes the network slightly more robust to the location of features within an image.

3.5.3 The VGG network

The VGG network was developed by the Visual Geometry Group in Oxford, it consists of repeating blocks of convolutional layers followed by three densely connected layers (Simonyan & Zisserman, 2015). This network architecture was developed using the ImageNet dataset

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 3.10 VGG architecture (Simonyan and Zisserman, 2015). Each network is broken into convolutional blocks and the classification stages are made up of three fully connected layers. The final fully connected layer has 1000 neurons, one for each class in the ImageNet dataset. All hidden layers use the ReLU activation function, the final layer uses a softmax activation function.

(Deng et al., 2009). The convolutional weights used in VGG net are all 3x3 (three samples by three samples) filters, this was a novel architectural selection at the time of development. Due to the sparse nature of the convolutional layer weights, and the dense nature of the fully connected layer weights, the majority of the parameters in the VGG net are in the fully convolutional portion of the network. Figure 3.10 shows the different forms of VGG architecture, the most commonly used architectures are the 16-layer (D) and 19-layer (E).

3.5.4 The backpropagation algorithm

The backpropagation algorithm gives deep neural networks their capacity to learn from labelled training examples. The understanding behind backpropagation is that by adjusting weights in the hidden layers, with respect to an error function, useful internal features are created (Rumelhart et al., 1986). The first step in backpropagation is to generate a network predicted output $\hat{y}_p$ based on the input vector for a particular input-output pair (x_p, y_p). The aim of backpropagation is to adjust the network weights to reduce the difference between the predicted

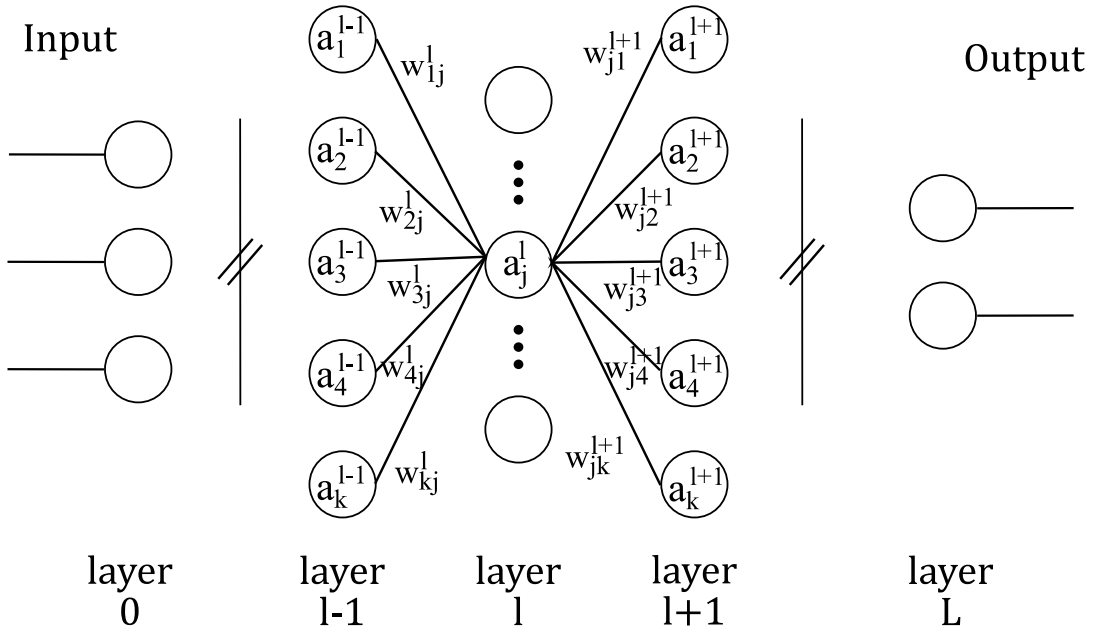


Figure 3.11 A feedforward neural network example with weights and node values with reference to a node in a hidden layer. The influence of biases has been omitted.

output $\hat{y}_p$ and the corresponding target output y_p, this is encoded in the loss function. The rule for changing network weights is

$$w_{ij}^{l} \leftarrow w_{ij}^{l} + \Delta w_{ij}^{l} \tag{3.16}$$

where

$$\Delta w_{ij}^{l} = -\eta \frac{\partial E}{\partial w_{ij}^{l}} \tag{3.17}$$

E represents the calculated loss based on the input-output pairs. Loss functions such as squared error, Equation 3.13, binary cross-entropy, Equation 3.14, or categorical cross-entropy, Equation 3.15, can be utilised to calculate E.

The following sections will focus on the loss value for one training pair, the p^{th}pair, this loss is E_p. The loss over the full set of training pairs is

$$\mathrm{E} = \sum_{p} E_p \tag{3.18}$$

Further, the gradients of loss with respect to each weight for all training pairs can be combined by

$$\frac{\partial E}{\partial w_{ij}^{l}} = \sum_{p} \frac{\partial E_p}{\partial w_{ij}^{l}} \tag{3.19}$$

The weight update for a training pair can be defined as

$$\Delta_p w_{ij}^l = -\eta \frac{\partial E_p}{\partial w_{ij}^l} \tag{3.20}$$

The aim of the backpropagation algorithm is to calculate these weight updates for each trainable parameter in the network. This can be achieved using the chain rule

$$\frac{\partial E_p}{\partial w_{ij}^l} = \frac{\partial E_p}{\partial a_j^l} \frac{\partial a_j^l}{\partial w_{ij}^l} \tag{3.21}$$

From Figure 3.7 and combining Equation 3.11 and Equation 3.12, it is known that

$$a_j^l = f\left(z_j^l\right) = f\left(\sum_i w_{ij}^l a_i^{l-1} + b_j^l\right) \tag{3.22}$$

hence,

$$\frac{\partial a_j^l}{\partial w_{ij}^l} = \frac{\partial a_j^l}{\partial z_j^l} \frac{\partial z_j^l}{\partial w_{ij}^l} \tag{3.23}$$

$$\frac{\partial z_j^l}{\partial w_{ij}^l} = a_i^{l-1} \tag{3.24}$$

$$\frac{\partial a_j^l}{\partial z_j^l} = \frac{\partial f\left(z_j^l\right)}{\partial z_j^l} \tag{3.25}$$

Equation 3.25 is the derivative of the activation function; the most common activations and their derivatives are illustrated in Table 3.3. This activation derivative Equation 3.23 can be summarised as

$$\frac{\partial a_j^l}{\partial w_{ij}^l} = a_i^{l-1} g_{pj}^l \tag{3.26}$$

where g_{pj}^l is the derivative of the activation function for the pth training pair.

Now the error with respect to each weight can be rewritten as

$$\frac{\partial E_p}{\partial w_{ij}^l} = \frac{\partial E_p}{\partial a_j^l} a_i^{l-1} g_{pj}^l \tag{3.27}$$

From Figure 3.11

$$\frac{\partial E_p}{\partial a_j^l} = \sum_k \frac{\partial E_p}{\partial a_k^{l+1}} \frac{\partial a_k^{l+1}}{\partial a_j^l} \tag{3.28}$$

now

$$\frac{\partial a_k^{l+1}}{\partial a_j^l} = w_{jk}^{l+1} g_{pk}^{l+1} \tag{3.29}$$

therefore

$$\frac{\partial E_p}{\partial a_j^l} = \sum_k \frac{\partial E_p}{\partial a_k^{l+1}} w_{jk}^{l+1} g_{pk}^{l+1} \tag{3.30}$$

The loss in the final network layer, layer L, can be easily calculated. So, $\frac{\partial E_p}{\partial a_k^L}$ for each output neuron can be worked out for each input-output training pair. Using Equation 3.30 the $\frac{\partial E_p}{\partial a_j^l}$ for each previous layer can be generated for each neuron. This must be repeated for the entire network, until the influence of loss on the weights in the first layer is calculated. Equation 3.27 can then be utilised to calculate the gradient of the loss with respect to each weight parameter. The combination of these losses over a set of training pairs can then be used to update the hidden weights of the network. The weight updates for the bias parameters are calculated in a similar manner.

The commonly applied mini-batch stochastic gradient descent (SGD) algorithm updates the internal weights based on the accumulated gradients of mini-batches (Howard & Gugger, 2020). The training examples for each mini-batch are randomly chosen from the overall set of training pairs; the m^{th} training pairs in each batch. In the SGD algorithm, mini-batches of size N_m are randomly chosen until all of the pairs in the training set have been exhausted. Once all the training pairs have been selected through the mini-batch selection routine, an epoch of training has been completed. The cost function η is scaled based on the size of the mini-batch, N_m, giving the new update rule,

$$w_{ij}^l \leftarrow w_{ij}^l + \Delta_m w_{ij}^l. \tag{3.31}$$

$$\Delta_m w_{ij}^l = \frac{-\eta}{N_m} \sum_m \frac{\partial E_m}{\partial w_{ij}^l} \tag{3.32}$$

Utilising mini-batches during training results in a smoother training loss over successive epochs. Batch sizes are often limited by the available hardware; larger batch sizes mean that more calculations need to be completed simultaneously.

The learning rate, η, is a small positive value which controls the size of each weight update step. Smaller learning rates result in slower network training as the trainable network parameters are only changed by a small amount after each mini-batch. Very large learning rates

may cause the weight updates to overshoot the optimal value, potentially causing the loss value to oscillate between sub-optimal values or even to diverge.

Momentum, μ, can also be incorporated into the weight update rule (Sutskever et al., 2013). Momentum is calculated using the gradient history, by accounting for the change in the learning gradient. Nesterov momentum is utilised in this work, which intuitively first moves in the direction of previous momentum and then corrects based on the new weight update gradient. This helps the algorithm to avoid some local minima in the non-convex surface of the cost function. A velocity term v is introduced for each weight in the network, the gradients will now change the velocity, rather than changing the position directly. In classical momentum, the velocity term for a particular hidden layer weight is

$$v_{ij}^{l} \leftarrow \mu v_{ij}^{l} - \eta \Delta w_{ij}^{l}. \tag{3.33}$$

For Nesterov momentum the gradient of weights is calculated after the initial step of μv_{ij}^{l} has been applied. The new weight update rule, which includes the newly updated velocity is

$$w_{ij}^{l} \leftarrow w_{ij}^{l} + v_{ij}^{l}. \tag{3.34}$$

By including momentum in the gradient descent algorithm, the training loss should move in a more consistent direction on each epoch, without large deviations because each step is not independent of previous steps.

3.5.5 Data representation: alternatives to feature extraction

The chosen representation for data is an important aspect of supervised learning in both shallow and deep learning algorithms. If important distinguishing information is not maintained in the chosen data transformation, then the algorithm will not be capable of learning to differentiate between classes. If too much irrelevant information is in the chosen representation this makes the classifier's task more challenging.

Historically EEG signals have been represented by sets of features which describe the signal characteristics in the temporal and frequency domains. These features usually describe windows of the temporal EEG; by shifting the windows a time-varying representation of the signal is produced. Features are chosen to be informative and to represent a compressed form of the input data which is relevant to the desired task. This has the effect of reducing the required computational complexity of the machine learning model being utilised and highlighting the important characteristics in the data. Figure 3.12 presents three different representations of an 80 second EEG signal. Each signal representation method demonstrates

the signal characteristics in 8-second sliding windows with no-overlap between windows. The first row shows the raw EEG signal. The second row in Figure 3.12 shows the signal represented as temporal windows where the only transformation is a downsampling operation.

The time-frequency representation of a signal is typically a visual representation of the frequency components as a function of time. An image is created using the time and frequency as axes, the colour intensity represents energy density for each combination of time and frequency. A common implementation of the time-frequency representation is the spectrogram.

To generate a spectrogram the magnitude of the frequency spectrum is taken across moving windows of a signal. These one dimensional outputs are then combined by ordering them in time to create a two dimensional representation. The windows can be overlapped to produce a smoother image, larger windows mean that more of the frequency axis is represented. The information about the phase from each frequency calculation is lost when creating a spectrogram. Row three in Figure 3.12 shows a time-frequency representation of the signal. The fourth row of Figure 3.12 shows the same windows of EEG represented as a set of 10 features.

3.6 Conclusion

This chapter has presented a brief history of automated seizure detection algorithms. Early algorithms were purely feature driven; bespoke features were designed using medical knowledge and empirical thresholds applied to categorise seizures in EEG. Further development and the search for more complex features was the next iteration of this feature driven approach. The inclusion of machine learning based algorithms trained with large sets of data-driven features resulted in more robust classifiers. The ANSeR algorithm, an SVM based algorithm which is combined with a large feature-set of relatively simple features, has finished clinical trials; this represents the current state-of-the-art in neonatal seizure detection.

Deep learning algorithms have the potential to overcome the limitations imposed by machine learning algorithms. Deep learning algorithms do not require an intermediate hand-engineered feature set, unlike machine learning these algorithms can learn from unstructured data. The inference stage of deep learning algorithms requires relatively simple mathematical operations, making these algorithms suitable for resource limited embedded systems implementations. This chapter outlined the MLP and CNN classifiers and the algorithms used to train them.

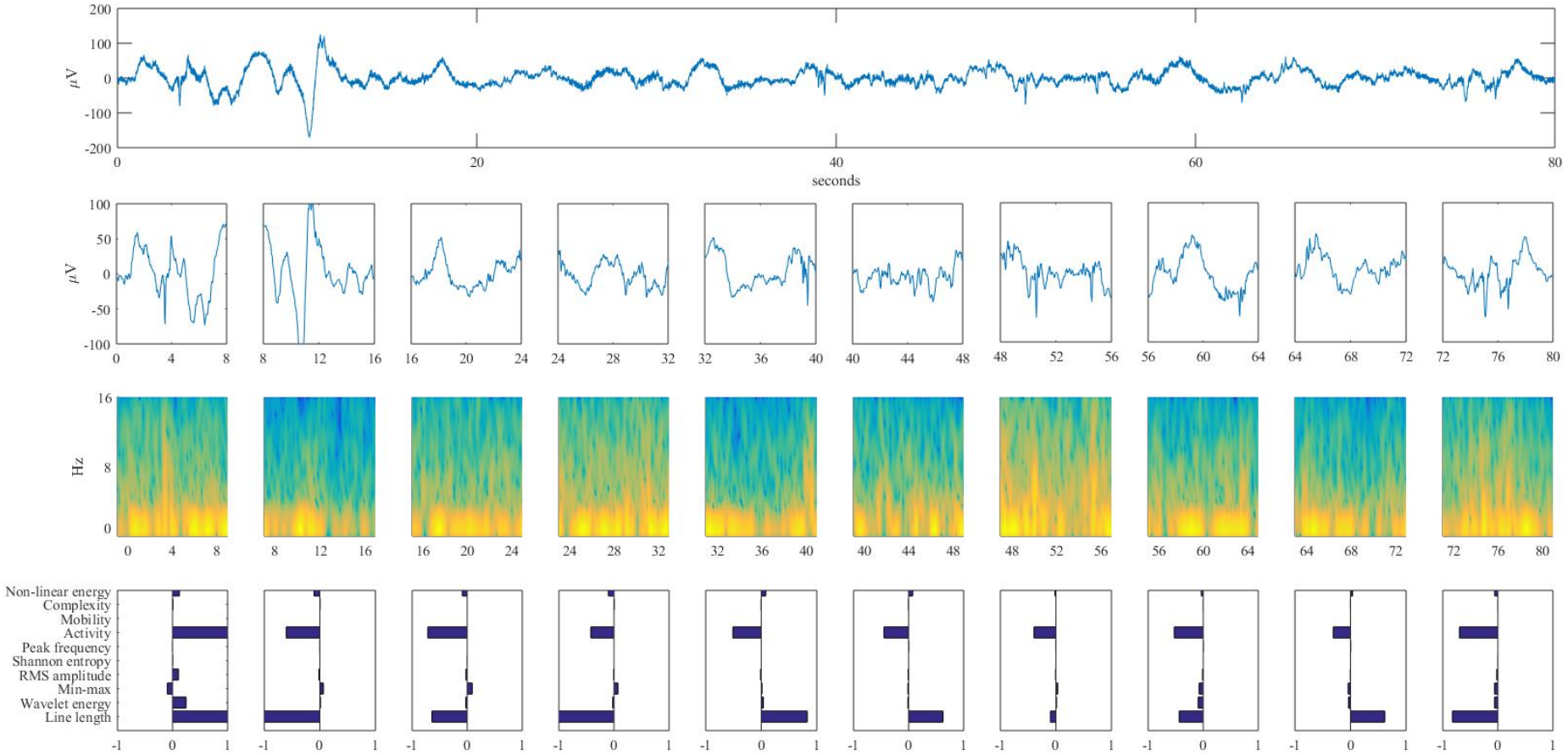


Figure 3.12 One channel of EEG split into 8-second windows. Each 8-second window is represented in three ways; as a temporal signal, a spectrogram and a set of 10 features {Non-linear energy, Complexity, Mobility, Activity, Peak frequency, Shannon entropy, RMS amplitude, Min-max, Wavelet energy, Line length}.

3.7 References

Aarabi, A., Wallois, F., & Grebe, R. (2006). Automated neonatal seizure detection: A multistage classification system through feature selection based on relevance and redundancy analysis. *Clinical Neurophysiology*, *117*(2), 328–340.

Ahmed, R., Temko, A., Marnane, W. P., Boylan, G., & Lightbody, G. (2017). Exploring temporal information in neonatal seizures using a dynamic time warping based SVM kernel. *Computers in Biology and Medicine*, *82*, 100–110.

Ansari, A. H., Cherian, P. J., Dereymaeker, A., Matic, V., Jansen, K., De Wispelaere, L., Dielman, C., Vervisch, J., Swarte, R. M., Govaert, P., Naulaers, G., De Vos, M., & Van Huffel, S. (2016). Improved multi-stage neonatal seizure detection using a heuristic classifier and a data-driven post-processor. *Clinical Neurophysiology*, *127*(9), 3014–3024.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–152.

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

Clancy, R. R., & Legido, A. (1987). The Exact Ictal and Interictal Duration of Electroencephalographic Neonatal Seizures. *Epilepsia*, *28*(5), 537–541.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

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

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

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

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, *3*(Mar), 1157–1182.

Howard, J., & Gugger, S. (2020). Chapter 4: MNIST basics. *In Deep Learning for Coders with fastai and PyTorch*. O'Reilly.

Kitayama, M., Otsubo, H., Parvez, S., Lodha, A., Ying, E., Parvez, B., Ishii, R., Mizuno-Matsumoto, Y., Zoroofi, R. A., & Snead, O. C. (2003). Wavelet analysis for neonatal electroencephalographic seizures. *Pediatric Neurology*, *29*(4), 326–333.

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

Mathieson, S., Rennie, J., Livingstone, V., Temko, A., Low, E., Pressler, R. M., & Boylan, G. B. (2016). In-depth performance analysis of an EEG based neonatal seizure detection algorithm. *Clinical Neurophysiology*, *127*(5), 2246–2256.

Ng, A. Y., & Jordan, M. I. (2002). On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. *Advances in Neural Information Processing Systems 14*, 841–848.

Optimize a Cross-Validated SVM Classifier Using bayesopt—MATLAB & Simulink. (2021). Retrieved 2 April 2021, from https://www.mathworks.com/help/stats/bayesian-optimization-case-study.html.

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

Platt, J. C. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*, 61–74.

Qu, H., & Gotman, J. (1997). A patient-specific algorithm for the detection of seizure onset in long-term EEG monitoring: Possible use as a warning device. *IEEE Transactions on Biomedical Engineering*, *44*(2), 115–122.

Ramantani, G., Schmitt, B., Plecko, B., Pressler, R. M., Wohlrab, G., Klebermass-Schrehof, K., Hagmann, C., Pisani, F., & Boylan, G. B. (2019). Neonatal Seizures—Are We there Yet? *Neuropediatrics*, *50*(5), 280–293.

Roessgen, M., Zoubir, A. M., & Boashash, B. (1998). Seizure detection of newborn EEG using a model-based approach. *IEEE Transactions on Biomedical Engineering*, *45*(6), 673–685.

Rumelhart, D. E., Hintont, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv:1409.1556 [Cs]*.

Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the importance of initialization and momentum in deep learning. *Proceedings of the 30th International Conference on Machine Learning (ICML)*, III–1139–III–1147.

Tapani, K. T., Vanhatalo, S., & Stevenson, N. J. (2019). Time-Varying EEG Correlations Improve Automated Neonatal Seizure Detection. *International Journal of Neural Systems*, *29*(04), 1850030.

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

Temko, A., & Lightbody, G. (2016). Detecting Neonatal Seizures With Computer Algorithms. *Journal of Clinical Neurophysiology*, *33*(5), 394–402.

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

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

Thomas, E. M., Temko, A., Lightbody, G., Marnane, W. P., & Boylan, G. B. (2010). Gaussian mixture models for classification of neonatal seizures using EEG. *Physiological Measurement*, *31*(7), 1047–1064.

Thomas, E. M., Temko, A., Marnane, W. P., Boylan, G. B., & Lightbody, G. (2013). Discriminative and Generative Classification Techniques Applied to Automated Neonatal Seizure Detection. *IEEE Journal of Biomedical and Health Informatics*, *17*(2), 297–304.

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., & Vapnik, V. (2000). Feature Selection for SVMs. *Advances in Neural Information Processing Systems*, 668–674.

Chapter 4: Developing a deep learning classifier

4.1 Introduction

In this chapter, the performance of deep learning architectures for various EEG input representations is presented. Experiments to develop an end-to-end deep learning classifier for neonatal seizure detection without relying on engineered features are reported. The different classification architectures which can be employed as part of a deep convolutional network are outlined. Details about the frameworks utilised in this thesis to implement the deep learning algorithms from a hardware and a software perspective are reported.

The utilization of deep learning represents a new approach in neonatal EEG classification. The hierarchical learning capabilities of deep architectures allow engineers to design algorithms which extract informative characteristics from raw inputs through an end-to-end optimisation procedure, circumventing the requirement for hand-crafted feature engineering. Previous neural network approaches to neonatal and adult seizure detection algorithms have combined the multi-layer perceptron (MLP) with engineered feature sets (Aarabi et al., 2006; Webber et al., 1996). In contrast to these previous investigations, where deep learning is applied to engineered features, the aim of this work is to learn from minimally preprocessed EEG inputs, this reduces the opportunity for human induced biases in the algorithm development process.

The convolutional neural network (CNN) is a deep learning algorithm which was originally developed for image processing applications. Due to the success of CNN algorithms in the ImageNet challenge (J. Deng et al., 2009; Krizhevsky et al., 2012; Simonyan & Zisserman, 2015) they have become widely used in varying applications. A review of deep learning for EEG analysis highlighted that the number of research publications which utilise CNN based architectures for various types of EEG processing has steadily increased since 2014 (Roy et al., 2019).

An important aspect to consider while designing a CNN architecture is the selection of an appropriate input representation. Different EEG representations may affect the performance of the deep learning algorithm, and the appropriate EEG representation to choose might be task dependent. This chapter presents the initial investigations into the development of a deep learning neonatal seizure detection algorithm with various methods of representing neonatal EEG as an input.

4.2 Review of input EEG representations for deep learning systems

A good starting point to consider when developing a seizure detection algorithm for neonates is to investigate commonly used approaches in the domain of seizure detection in adults. Deep learning has been utilised in a wide range of adult EEG labelling tasks including seizure prediction and detection, sleep stage classification as well as cognitive and affect monitoring (Perslev et al., 2019; Thodoroff et al., 2016; Ullah et al., 2018; Zheng & Lu, 2015). The recent increase in publications utilizing deep learning for EEG classification is motivated by the benefits of automatic feature learning. The prevalence of epilepsy in adults has motivated research into automated seizure detection and prediction algorithms. In recent years, many of the developed algorithms rely on deep learning (Acharya et al., 2018; Ozcan & Erturk, 2019; Truong et al., 2018). Many studies skip the traditional step of feature extraction and some additionally omit any artifact rejection stage, placing the burden of learning from potentially noisy training data on the deep neural network (Roy et al., 2019).

As previously mentioned, the CNN deep learning architecture was initially developed for image processing applications. One possible approach to develop a neonatal seizure detection algorithm is to apply a CNN to a suitable image representation of EEG. There are a variety of different ways to convert EEG to an image; any transformation which represents the original EEG signal in 2-dimensions could become the input to a 2-dimensional CNN algorithm.

One way to generate images from multi-channel EEG signals, is to project the positions of EEG electrodes on the scalp onto a 2-dimensional space. The placement of electrodes is in 3-dimensions over the scalp of the newborn, therefore careful consideration must be taken during the projection from 3-dimensions to 2-dimensions to preserve the inter-electrode distances and dependencies. A 2-dimensional mapping of features from individual EEG channels into images was proposed by Ozcan *et al.* (Ozcan & Erturk, 2019). In this representation, the x and y dimensions represented the position on the scalp and the z dimension showed the value of a feature at that point. Each feature was mapped in 3-dimensions in this fashion. This approach still relied on the decomposition of EEG windows into features. The features utilised were the power in five spectral bands, the first four statistical moments (mean, variance, skewness and kurtosis) and two of the Hjorth parameters (mobility and complexity). These features are known to be informative indicators of seizures in adult EEG.

Another alternative 2-dimensional transformation could be based on time-frequency analysis. Utilising a Fourier decomposition based approach which shifts across time gives a 2-

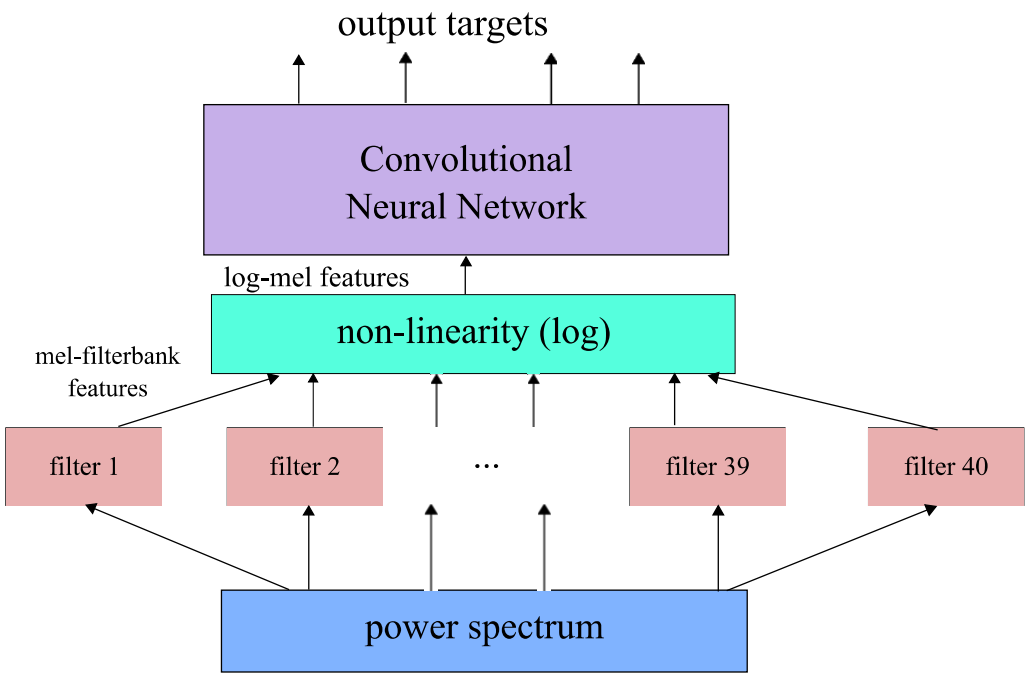


Figure 4.1 Block diagram of the learned filter-bank applied to spectrogram EEG inputs incorporated into the CNN architecture (Sainath et al., 2013). The weights in the filter bank layer each span a fixed set of samples across the frequency dimension of the 2-dimensional spectrogram input and these weights move across the time dimension to produce a weighted spectrogram output. A log operation is applied to this output. The filters are initialised with Mel-filter bank weights.

dimensional output from a 1-dimensional input signal. In this transformation the x and y dimensions represent time and frequency, and the z dimension is the power in that frequency at that time, resulting in a representation of the frequencies of a signal as they vary with time. Truong *et al.* designed a CNN based seizure prediction algorithm based on time-frequency representations of 16 channel adult EEG (Truong et al., 2018). The EEG representations were formatted into the time-frequency images described above by applying a Short-Time Fourier Transform (STFT) to 30-second windows of EEG. These 2-dimensional matrices are also known as spectrograms. A fourth input dimension represented the spectrograms across different EEG channels. The goal of creating a seizure prediction algorithm which generalises well across datasets with minimum feature engineering was achieved through deep learning.

In speech recognition, the transformation of temporal signals into 2-dimensional spectrograms through successive Fourier transform based operations is a widely used preprocessing step. In many previous works, the further processing of time-frequency representations using Mel-filter banks has improved performance of machine learning and deep learning classifiers (Hinton et al., 2012). The aim of these filters is to create a time-frequency representation where different frequency ranges are weighted based on a knowledge of how the human auditory system

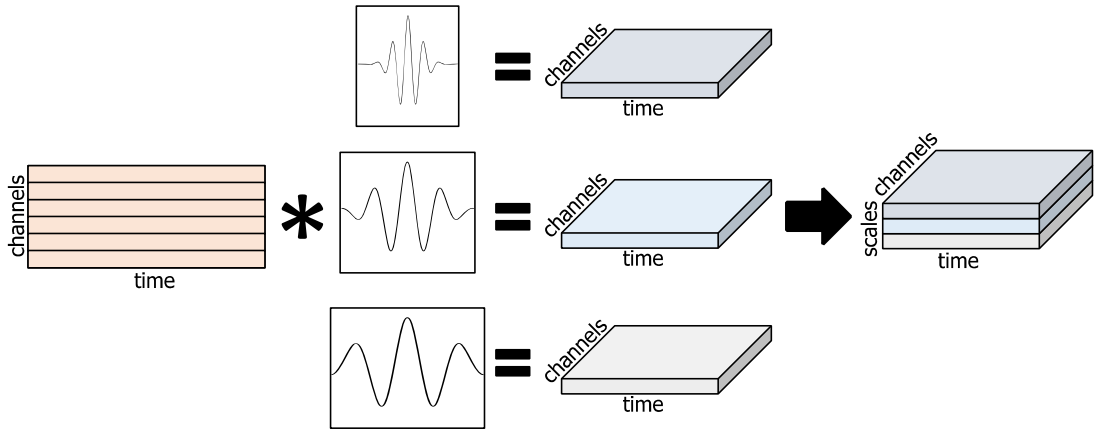


Figure 4.2 Time-frequency transformation of EEG utilising scaled wavelet filters. This image was published in (Khan et al., 2018).

functions. These Mel-filter banks are generic and are not optimised for different audio domain tasks; the rise of deep learning has allowed for the retraining of filter banks based on task specific objective functions (Qu et al., 2016; Sainath et al., 2013).

Both Sainath *et al.* (2013) and Qu *et al.* (2016) have utilised deep learning architectures to train custom filter banks as part of the end-to-end deep learning optimisation. In these algorithms, EEG inputs are transformed to spectrogram representations before being acted upon by a sparse locally connected filter bank layer. The filter bank spans a fixed set of samples in the frequency dimension and moves across the time dimension, this is reflective of how a filter in a convolutional layer moves across a layer input. A log operator is applied to the weighted outputs, this representation is then input to the first convolutional layer of a CNN. Figure 4.1 shows a block diagram of the sparse locally connected filter banks integrated into the CNN architecture. The output of this filter bank layer is a 2-dimensional time-frequency representation of the input signal, which is weighted based on the learned filter-banks.

A caveat to consider when applying frequency transformations to EEG signals is that EEG is a non-stationary signal. Neonatal EEG cannot be considered stationary, even on a short-time scale, for a multitude of reasons including the presence of spikes and other discrete events. The wavelet decomposition provides a representation of the frequency domain characteristics of the EEG signals and is more advantageous than Fourier based decompositions as wavelets are capable of accurately representing non-stationary signals, such as EEG (Kitayama et al., 2003; Temko & Lightbody, 2016).

In the study by Kahn *et al.* (2018) for epileptic seizure detection in adults, ten different scales of the Mexican-hat mother wavelet were applied continuously to multi-channel EEG resulting

in a data representation of the form (EEG channels, wavelet scale, time). A diagram of the multi-channel wavelet transformation is shown in Figure 4.2.

Roy *et al.* (2019) noted that there is an increase in the number of publications which combine raw temporal adult EEG signals with CNN classifiers; 49% of papers reviewed used only raw or minimally processed EEG signals as input. This trend can be seen in the increased number of epileptic EEG classification studies which employ no feature extraction or time-frequency transformations for seizure detection and prediction.

Acharya *et al.* (2018) developed a CNN architecture which learns to detect seizure events from raw temporal adult EEG waveforms, no feature extraction and no time-frequency transformations of the signal are required (Acharya et al., 2018). Their proposal of a 13-layer deep convolutional neural network was compared to studies which utilised traditional feature-based machine learning algorithms. Although the deep learning algorithm did not achieve the best performance, it reached a specificity of 90% and a sensitivity of 95% on the seizure detection task without the requirement for separate feature engineering or feature selection.

Deep convolutional neural networks have also been employed as feature extractors for neonatal seizure detection, where the input is multi-channel temporal EEG and the output of the algorithm is classified using a random forest (Ansari et al., 2018). The temporal EEG is represented as a 2-dimensional input where the x dimension shows the EEG signal in time and the y dimension represents the different EEG channels. This deep neural network and random forest hybrid architecture was shown to outperform feature based machine learning algorithms.

4.3 Review of deep learning EEG classification methods

Machine learning classification algorithms consist of a feature extraction stage and a classification stage. Deep learning algorithms for classification tasks also consist of a feature extraction stage and a classification stage. The difference between shallow and deep classification algorithms is that deep learning algorithms optimise the feature extraction stage and the classification stage in one end-to-end optimisation routine. The end-to-end optimisation typically involves weight updates based on the backpropagation algorithm, although other weight update methods can be employed, for example reinforcement learning (Zhao et al., 2017). There are many variables to consider when designing a convolutional neural network architecture for classification tasks; the structure of the classification stage is important and should be informed by domain-knowledge about the task and the format of the available labels.

The first CNN architecture employed stacked layers of dense neurons to classify inputs (LeCun et al., 1989). In the AlexNet architecture, the first CNN based algorithm to win the ImageNet challenge, the feature extraction portion of the network was implemented using convolutional layers and the classification utilised fully connected neural network layers. This architecture consisting of convolutional layers followed by fully connected dense layers is still very prevalent today (Alom et al., 2018; Simonyan & Zisserman, 2015).

An alternative CNN classification architecture is the fully convolutional network (FCN). In the FCN architecture, convolutional layers are used both to extract features from inputs and to perform classification. The FCN architecture offers numerous benefits when compared to a CNN that relies on fully connected layers. First, the number of network parameters is notably reduced, which, along with weight sharing, eliminates the need for regularisation routines such as dropout or L1-L2 regularisations. Second, the absence of a dense layer allows the network to classify inputs of any length, since the last global pooling layer can process a sequence of any length giving a constant sized output. Finally, yet importantly, since only convolutional layers are used, the temporal relationship between the input and the output is preserved. In computer vision, FCN architectures have been used for object detection and localisation tasks with state-of-the-art results (Dai et al., 2016; Yao et al., 2017).

A variation of this FCN architecture is the U-Net; this network utilises only convolutional layers to implement a contraction path followed by an expansion path (Long et al., 2015). The U-Net was developed for image segmentation tasks, where each pixel of an image should be labelled. The contraction path uses unpadded convolutions to generate a compressed image representation that captures pixel context, and the symmetric expansion path precisely maps this representation onto the original input feature space. A recent study applied a U-Net like architecture to the task of EEG sleep staging in adults (Perslev et al., 2019). This novel U-Time algorithm was able to outperform previous baselines on sleep stage classification benchmarks.

In the development of the U-Time architecture the results are compared to an implementation of a CNN-LSTM architecture. The Long Short-Term Memory (LSTM) is a specific form of Recurrent Neural Network (RNN) layer. RNNs represent a group of deep learning architectures which are designed for sequence modelling tasks. The internal states of an RNN architecture are dependent on previous states; this gives the network a mechanism to learn from the history of an input sequence. In the past, sequence modelling tasks were synonymous with RNN architectures. Results from recent studies have shown that CNN based architectures have the

ability to learn the temporal relationships which are important for sequence modelling tasks (Bai et al., 2018). The use of hybrid CNN-LSTM architectures has given promising results in EEG classification tasks (Craik et al., 2019). The addition of RNN layers is another possible approach to designing the classification stage of a CNN architecture.

All classification architectures mentioned here have involved the use of additional deep learning layers. Another possible classification architecture to consider when designing a deep learning algorithm is classifier stacking. Stacking is a machine learning technique where the outputs of a model become the inputs to another model. The models are trained separately, but it is assumed that the intermediate representation learned by the first model will be a set of informative features that can be used to train the second model. In a recent publication, a neonatal seizure detector was designed utilizing a CNN as a feature extraction stage followed by a decision tree as a classifier (Ansari et al., 2018). In this algorithm, the CNN is a complex feature extraction routine and the decision tree is a machine-learning algorithm, which classifies based on the extracted features.

While some of the options for implementing the classification stage of a CNN architecture have been outlined, there are still many possible deep learning architectures which could be exploited to develop a neonatal seizure detection algorithm. Architecture design and selection are an important part of the deep learning architecture development process.

4.4 Evaluation of neonatal EEG representations for classification

In this section, early results from deep learning experiments on Term Dataset I will be illustrated. Term Dataset I contains EEG signals and seizure annotations; more details are available in Chapter 2, Section 2.6.1. These initial experiments were designed to examine if deep learning can extract meaningful features from unprocessed EEG signals. The application of deep learning to neonatal EEG is a relatively new research area so there is no clear prior research that suggests the best representation for neonatal EEG signals. These experiments were also conducted to help determine the suitability of different EEG representations.

The preprocessing procedure described for the Cork SVM algorithm is applied to the raw EEG signals; see Chapter 3, Section 3.3.2. The data is split into 8-second windows with a 1-second shift between successive windows. Due to the need for per-channel annotations only EEG windows with channel-specific seizure annotations are in the seizure class, a sample of the non-seizure windows are utilised during training to balance the data classes. Data from twelve of

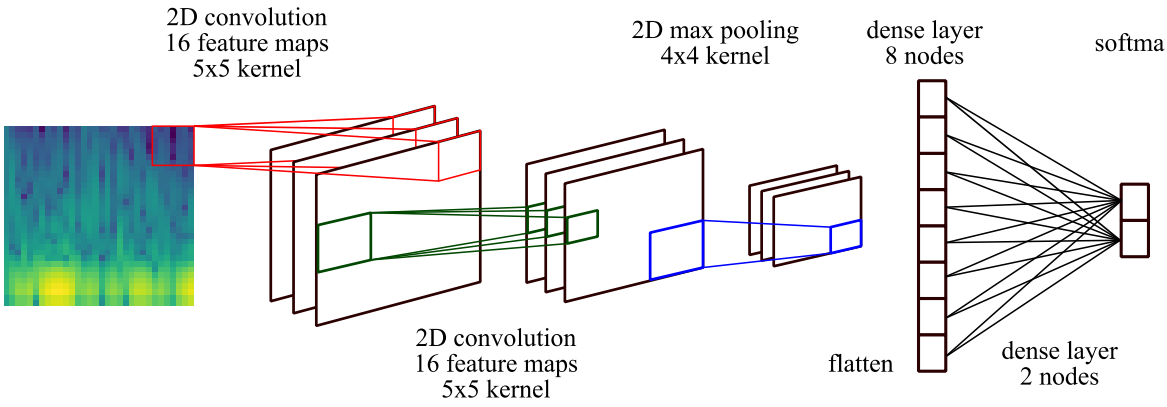


Figure 4.3 A CNN architecture for processing 2-dimensional EEG spectrograms and predicting seizure or non-seizure labels. Each spectrogram is 33x33 samples and represents an 8-second window of EEG sampled at 32Hz. The convolutional layers have 3-dimensions; the height and width of the input image and the feature map dimension. The dense layers require a 1-dimensional input so a flatten layer is utilised to transform the 3-dimensional representation into a 1-dimensional representation.

the infants are used for algorithm training, six are used for algorithm testing. Performances are reported as AUC values in percentages.

4.4.1 Spectrogram representation of EEG

To investigate the performance of an automated seizure detection algorithm with spectrogram input representations of EEG, an experiment was conducted using Term Dataset I. The EEG samples from twelve infants were transformed into spectrograms by applying 64 sample fast Fourier transform (FFT) with an overlap of 8 samples to each 8-second window thus leading to a 33x33 image for each window of EEG. A CNN architecture with two convolutional layers

Table 4.1. A CNN architecture for processing spectral inputs which was utilised in experiments. The convolutional layers are designed to act as feature extractors and the densely connected layers should learn to classify based on the learned intermediate representation. The shape of the output after each network layer is shown and the number of trainable parameters in each layer is also presented.

Layer Type	Activation	Shape	Output Shape	Parameters
Input		33x33	1x33x33	0
2D Convolution	ReLU	16 filters 5x5 kernel stride 2x2	16x17x17	416
2D Convolution	ReLU	16 filters 5x5 kernel stride 2x2	16x9x9	6416
Max Pooling		pool 4x4 Stride 2x2	16x3x3	0
Flatten			144	0
Dense	ReLU	8 nodes	8	1160
Dense	Softmax	2 nodes	2	18

Figure 4.4 The performance of the network on a test dataset plotted against the number of training iterations. The EEG representation in this network was an FFT based spectrogram 2-dimensional input. The red dashed line indicates the best performance on the test set. The CNN architecture is shown in Figure 4.3 and Table 4.1.

followed by two fully connected dense layers was designed, this is shown in more detail in Figure 4.3 and Table 4.1. A categorical cross-entropy cost function and mini-batch stochastic gradient descent were utilised during network training with a learning rate of 0.01. The learning rate was reduced by 10% every 20 epochs; the batch size was 128.

The convolutional layers each had 16 filters each of size 5x5. These moved across the layer inputs with a stride of 2 in the vertical and horizontal direction. A max pooling layer with a filter size of 4x4 and a 2x2 stride reduced the size of the learned representation. This 3-dimensional representation was then flattened. The output of the flatten layer was input to a fully connected layer with 8 nodes. Then a final fully connected layer with 2 nodes followed by a ReLU activation classified the input as either (0, 1) or (1, 0); this represented the seizure and non-seizure classes, respectively.

The performance of this architecture was reported on the six infants in the test set after each training epoch. Figure 4.4 shows the performance of the trained model as a function of the training epochs. The best achieved performance on the test dataset was 85% AUC. This performance indicated that the model learned to detect seizures, but the performance does not compare favourably with that of previous state-of-the-art algorithms.

4.4.2 Frequency-scaled spectrogram representation of EEG

When interpreting EEG for medical and engineering decision-making the signal is often split into frequency sub-bands; these are usually the delta, theta, alpha, and beta frequency bands which were outlined in Chapter 2 Section 2.3. To transform the spectrogram which was used in experiments in Section 4.4.1, an engineered filter bank could be applied to emphasise the characteristics of different frequency ranges. Previous machine learning algorithms have used the characteristics of different frequency bands within an EEG window to distinguish neonatal seizures. A deep learning framework could be utilised to learn a set of task specific filter bank weights through the backpropagation algorithm. This could give a more nuanced representation of the EEG signal as a set of learned frequency bands. The problem of detecting neonatal

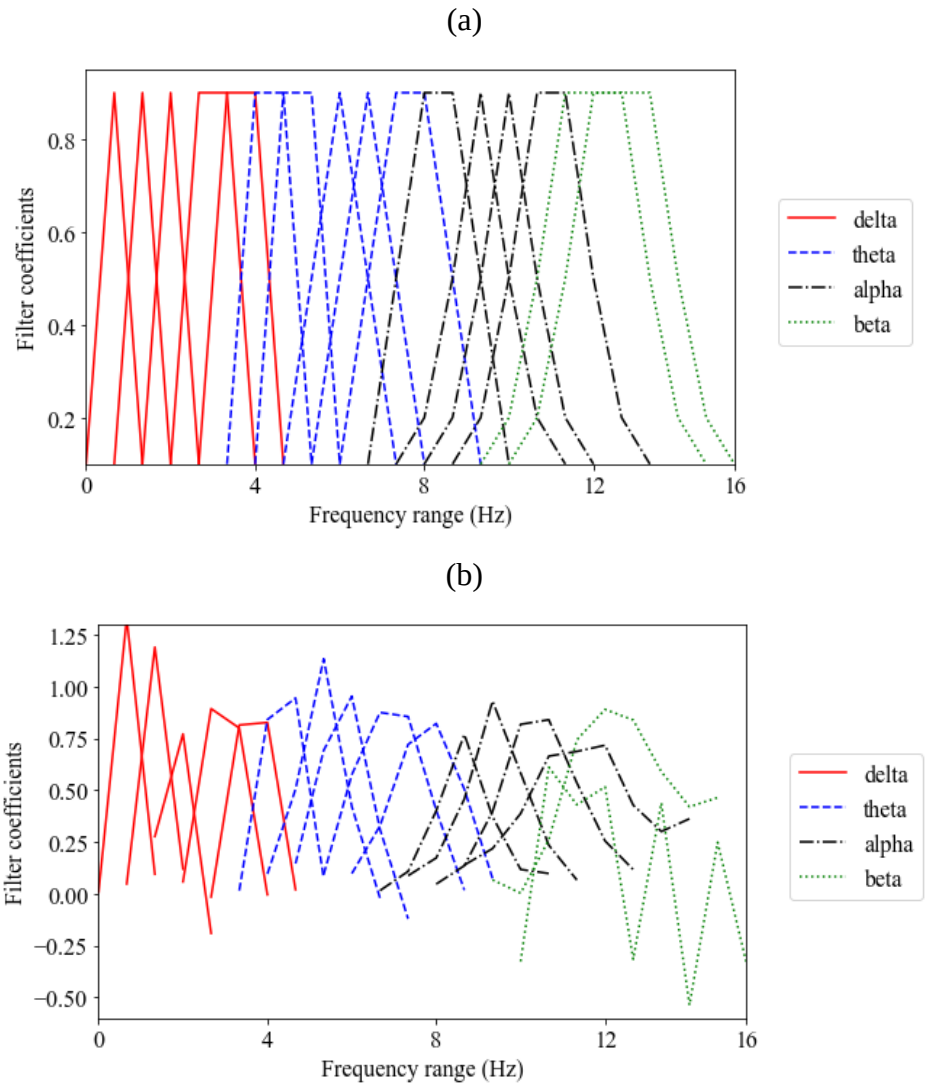


Figure 4.5 The initialised filter values are represented in (a). The filter values after being trained as part of the backpropagation routine are represented in (b). The different line styles approximately map each filter to the frequency bands of EEG.

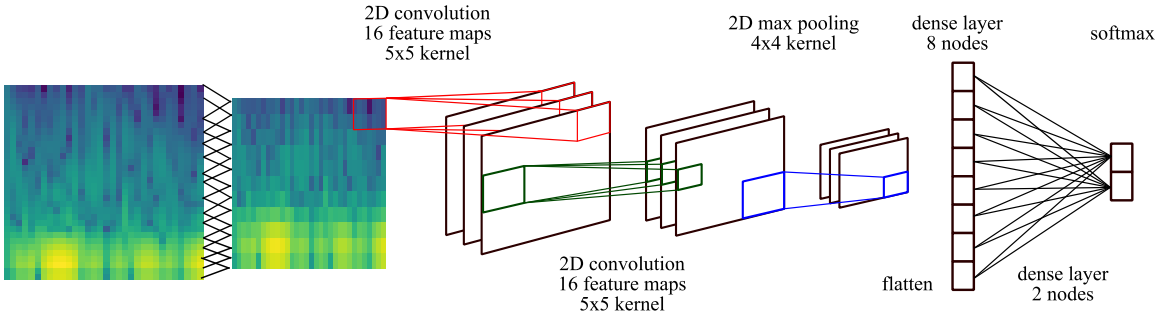


Figure 4.6 CNN architecture for processing EEG spectrograms, the first layer of this network is a weighted series of filter banks which can be frozen or learned by the network.

seizures from EEG windows can be reformulated as one of learning the filter-bank weights of the spectral envelope.

The concept of representing a signal as a weighted transform of the signal frequencies is similar to Mel-scaling in audio processing. To achieve this, overlapping triangular filters were defined, these filters map the higher resolution magnitude spectrum to a matrix which represents local frequency bands across time. The initialised filter bank is illustrated in Figure 4.5 (a), filters in the lower frequency range span fewer bins and at higher frequencies there are fewer filters and these span wider frequency ranges; this was designed to emphasise lower frequencies which are known to be more discriminative when detecting neonatal seizures. Each filter is applied to a fixed set of frequency bins, but all filters shift across the time dimension of the spectrogram with a stride of one sample; this striding behaviour is similar to a traditional convolutional filter. At the output of this layer EEG windows are represented with the scaled frequency in the y dimension and the x dimension represents time.

Figure 4.6 shows a CNN architecture where the first layer consists of sparsely connected weighted multiplications, these act as a trainable filter bank layer. The subsequent layers of the network remain unchanged from the network in Figure 4.6 and Table 4.1. This frequency-scaled network was trained under the same conditions utilised in Section 4.2.1 of this chapter; the same training and testing datasets, and the same optimization hyperparameters were employed.

A categorical cross-entropy cost function and mini-batch stochastic gradient descent were utilised during network training with a learning rate of 0.01. The learning rate was reduced by 10% every 20 epochs; the batch size was 128. The performance on the test dataset is shown in Figure 4.7. In one experiment, the filter banks were static and untrainable. In a second

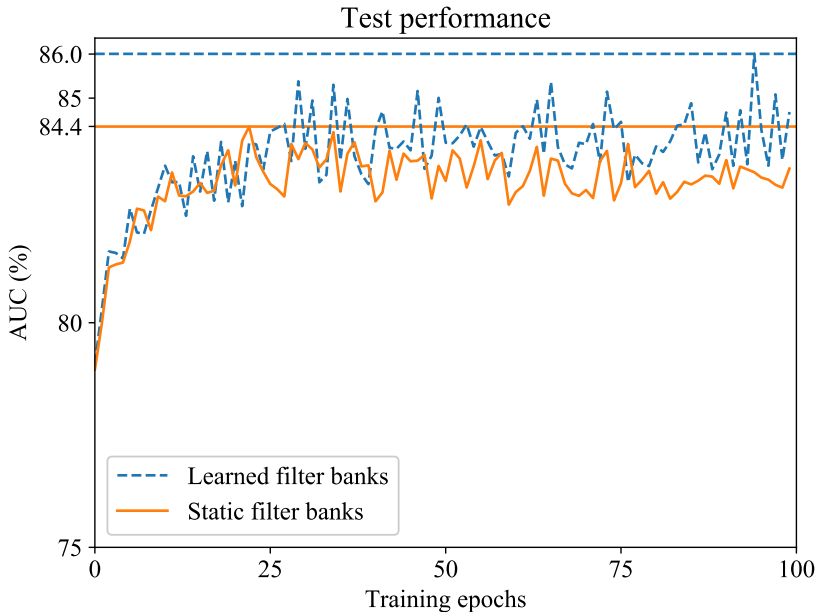


Figure 4.7 The EEG representation in this network was an FFT based spectrogram 2-dimensional input weighted by a filter bank. The solid orange line plots the performance on a test dataset in an experiment where the filters were frozen and untrainable; the solid horizontal orange line represents the best test performance from this experiment. The dashed blue line plots the performance on a test dataset in an experiment where the filters were learned during the deep learning training; the dashed horizontal blue line represents the best test performance from this experiment.

experiment, the filter banks were trainable and the weights were adjusted based on the gradients learned through backpropagation. Figure 4.5 (b) represents the filter bank values after the experiment when they were learned through the backpropagation algorithm.

4.4.3 Temporal representation of EEG

Untransformed temporal EEG signals can also be utilised as input to CNN architectures. This requires a change in the CNN architecture; 2-dimensional filters are replaced with 1-dimensional filters which shift along the time axis of the input EEG signal. To assess the

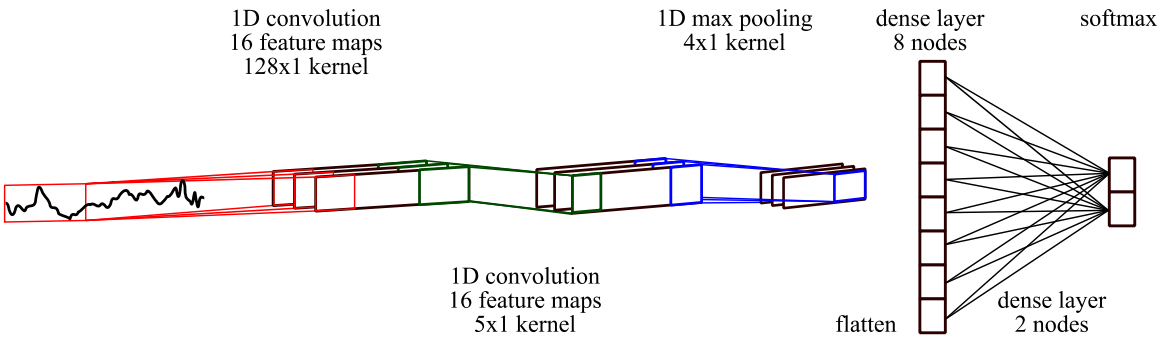


Figure 4.8 CNN architecture for processing temporal EEG signals. All of the convolutional and pooling operations are implemented over 1-dimension. A fully connected layer is utilised in the classification stage after the feature map representation has been flattened to a single dimension.

Table 4.2. CNN architecture for processing temporal inputs.

Layer Type	Activation	Shape	Output Shape	Parameters
Input		256x1	1x256	0
1D Convolution	ReLU	16 filters 128x1 kernel stride 2x1	16x256	2064
1D Convolution	ReLU	16 filters 5x1 kernel stride 2x1	16x128	1296
1DConvolution	ReLU	16 filters 5x1 kernel stride 2x1	16x64	1296
Max Pooling		pool 4x4 Stride 2x2	16x31	0
Flatten			496	0
Dense	ReLU	8 nodes	8	3976
Dense	Softmax	2 nodes	2	18

relative performance of a CNN architecture which utilises raw temporal EEG signals as input the architecture in Figure 4.8 is utilised.

Table 4.2 shows the CNN architecture for processing temporal EEG. All the convolutional filters utilised in Table 4.1 are transformed to 1-dimensional filters. An extra layer is added to the top of the network to process the EEG. The filters in this layer are long filters, they correspond to 4-seconds of the input signal.

A categorical cross-entropy cost function and mini-batch stochastic gradient descent were utilised during network training with a learning rate of 0.01. The learning rate was reduced by

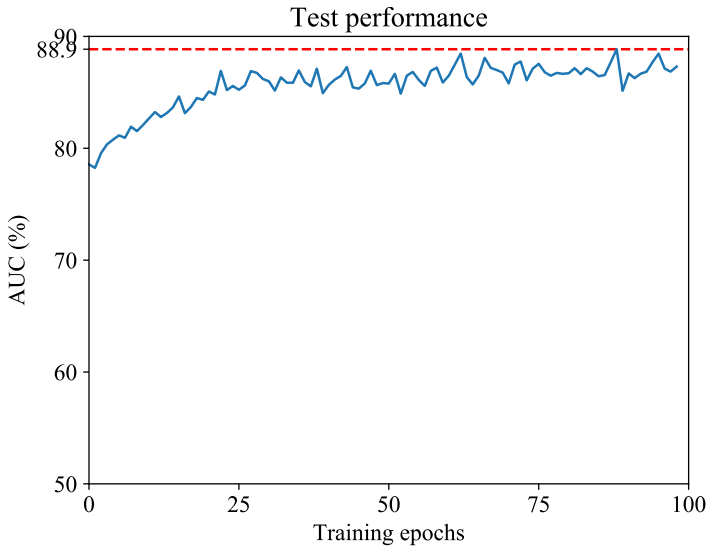


Figure 4.9 This learning curve shows the testset performance of the network with 1-dimensional temporal EEG as the network input against training epochs. The red dashed line indicates the best performance.

Table 4.3 Performance of experiments to evaluate the EEG representations for deep learning seizure detection architectures. The ANSeR baseline SVM algorithm results are also included in this table.

Representation	Architecture	Performance (AUC)
Feature set	SVM	96.6%
Spectrogram	CNN	85.0%
	CNN + static filter bank	84.4%
	CNN + learned filter bank	86.0%
Temporal	CNN	88.9%

10% every 20 epochs; the batch size was 128. Figure 4.9 shows the learning curve on the test dataset for the experiment utilizing temporal EEG as network input. The best performance achieved for this CNN architecture and temporal representation is 88.9% AUC.

4.4.4 Discussion of EEG representation experiments

The experiments described in this section are not exhaustive explorations of the possible architectures which could be defined to predict seizures from neonatal EEG representations. These experiments do, however, give an indication of the possible representations of EEG which could be exploited further to develop neonatal seizure detection algorithms.

Table 4.3 shows the AUC of four different experiments, three utilizing spectrogram inputs and one utilizing temporal inputs; the ANSeR baseline performance is also included as a reference point. It should be noted that the ANSeR algorithm performance that is reported here was tested on all infants in Cork Dataset I through leave one patient out (LOO) testing and so the performance cannot be compared to results of experiments in this chapter, but it is an interesting performance indicator. The same training and testing data were utilised in all four deep learning experiments and the network architectures were altered as minimally as possible to incorporate the different input formats.

When defining a feature set to represent EEG for the seizure detection task human induced biases influence the choice of features. Some of these human induced biases are also present when EEG is transformed with the STFT as the engineer is assuming that the signal is stationary. This mismatch between the *cos* and *sin* based decomposition and the presence of evolving characteristics and discrete events might be contributing to the results seen in Table

4.3, where the network with temporal inputs results in the best performance. This would imply that learning data-driven filters instead of using the cos/sin basis of the FFT is advantageous.

Figure 4.10 shows four different representations of neonatal EEG. Each row represents a different EEG event. In order for a neonatal seizure detection algorithm to learn, it needs to be able to differentiate between seizure events, non-seizure EEG and artifacts. Transforming signals can highlight important characteristics, but it can also introduce human biases as often some information is lost during the signal transformation. Clinical staff differentiate between these different event types from temporal representations of the signals; see the left-most column in Figure 4.10.

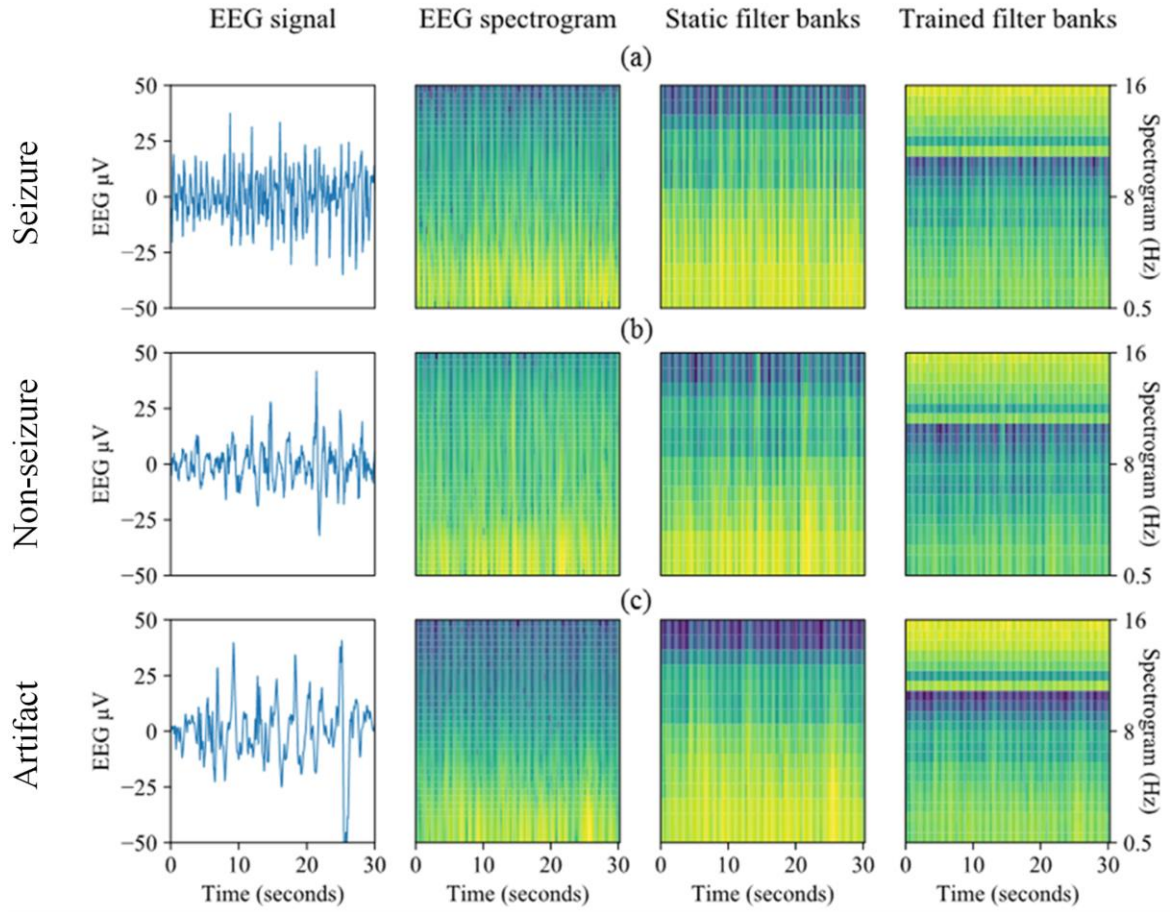


Figure 4.10 Three 30-second EEG segments represented as temporal signals and as images. The left-most column shows the EEG segments as preprocessed temporal signals. The left of centre column shows the spectrogram representation of the EEG signal. The right of centre and right-most columns show the spectrogram representations with static filter banks and filter banks learned through the backpropagation algorithm applied. The EEG in row (a) is from a seizure event. The EEG in row (b) is from a non-seizure period. The EEG in row (c) is from a segment of EEG marked as containing signal artifacts due to movement of the patient.

4.5 Framework for deep learning experiments

Training deep learning architectures is a computationally expensive process. Training is an iterative process where each iteration through the training data results in a discrete change to the network parameters; during training each iteration requires a forward pass operation and a backward pass operation. The forward pass results in predictions, the backward pass propagates the prediction errors to each trainable parameter in the network. Deep learning inputs are usually represented in higher dimensional spaces than are typically seen in other machine learning algorithms. Instead of a few bespoke features each input is represented by relatively large unprocessed input vector, matrix, or tensor. Each individual operation is computationally simple, a network mostly consists of successive addition and multiplication steps, but due to the number of network parameters, the iterative nature of training, and the high dimensional input space, numerous operations are required during network training and to a lesser extent during inference.

The central processing unit (CPU) is the primary processing unit on a computer. The CPU is a general purpose processor which can perform arithmetic operations and is also required to interface with other computer hardware components. Modern CPUs typically have a few cores, but they are not designed for high parallelization. A CPU's computational cores are intended for sequential instruction processing.

The graphical processing unit (GPU) is primarily intended for image rendering tasks on a computer's screen. A GPU has many cores which are suitable for running highly parallelised arithmetic operations and a GPU is also capable of efficiently fetching large inputs from memory. The utilisation of GPUs for training deep learning architectures has resulted in a paradigm shift, where larger models can be trained with more data in tractable timeframes (L. Deng, 2014). Increasingly deep learning architectures are being trained on GPUs in a parallelised fashion (Ben-Nun & Hoefler, 2018). The nature of mini-batch updates allows for the distribution of deep learning training across multiple cores in a GPU. The deep learning algorithms developed and tested in this work were processed using an NVIDIA V100 GPU.

In recent years application specific integrated circuits (ASIC) have been developed to meet the demands of deep learning training and inference. The most well-known of these is a tensor processing unit (TPU), which was developed by Google (Jouppi et al., 2017). These ASICs are further optimising the hardware for training deep learning algorithms through communications protocols, custom floating point formats and increased processing power.

Developments in software which can run deep learning models on parallel devices like GPUs have also been an important part of this deep learning paradigm shift. Many of these software libraries have been made available as open-access and this has enabled researchers across the world to train deep learning algorithms on GPUs (Abadi et al., 2016; Bergstra et al., 2011; Chollet, 2015). The algorithms in this work have been developed using Keras with both a Theano and a TensorFlow backend.

4.6 Conclusion

The initial experiments conducted in this chapter show that the performance of simple deep learning based seizure detection algorithms with off the shelf available architecture and EEG representations, including those that were borrowed from other areas like speech, fall short of the Cork SVM seizure detection algorithm baseline performance. During the development of Cork SVM many years of work were spent engineering and selecting the features. Deep learning has the advantage of learning from EEG representations which do not require laborious feature engineering routines. The engineering expertise and time that was previously spent designing feature sets can now be focused on the task of designing and optimising the deep learning architectures for the challenge of detecting seizures from neonatal EEG.

In 2015, the application of deep learning to neonatal seizure detection was not common practice. Initial assumptions about the performance of deep learning architecture in image processing domains encouraged the development of deep learning architecture for detecting neonatal seizure events which processed 2-dimensional time-frequency representations of EEG. Results from preliminary experiments showed that the performance of a simple neonatal seizure detection algorithm using spectral representations as input did not approach the baseline performance of the Cork SVM algorithm (96.6%). The application of both static and learned feature banks was explored in an effort to increase the performance of networks which learn to detect seizures from spectrogram representations of EEG.

An understanding of the non-stationary nature of neonatal EEG and domain expertise concerning how neonatal neurophysiologists label neonatal seizures, has prompted experiments where simple deep learning architectures were applied to temporal EEG signals. These results were closer to the baseline Cork SVM algorithm performance. Further architecture optimisation of deep learning algorithms for processing EEG as spectrograms is possible but the results in Table 4.3 motivate the use of temporal EEG signals as the input to a deep learning network.

Although the architectures explored in this chapter did not reach the performance of the Cork SVM baseline, the results show that with further development and engineering a neonatal seizure detection with deep learning is possible. In the next chapter, a novel deep learning architecture which was specifically designed to detect seizures in neonatal EEG is presented.

4.7 References

Aarabi, A., Wallois, F., & Grebe, R. (2006). Automated neonatal seizure detection: A multistage classification system through feature selection based on relevance and redundancy analysis. *Clinical Neurophysiology*, *117*(2), 328–340.

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Zheng, X. (2016). *TensorFlow: A System for Large-Scale Machine Learning*. 265–283.

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., & Adeli, H. (2018). Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Computers in Biology and Medicine*, *100*, 270–278.

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Van Esesn, B. C., Awwal, A. A. S., & Asari, V. K. (2018). The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches. *ArXiv:1803.01164 [Cs]*.

Ansari, A. H., Cherian, P. J., Caicedo, A., Naulaers, G., De Vos, M., & Van Huffel, S. (2018). Neonatal Seizure Detection Using Deep Convolutional Neural Networks. *International Journal of Neural Systems*, *29*(04), 1850011.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *ArXiv:1803.01271 [Cs]*.

Ben-Nun, T., & Hoefler, T. (2018). Demystifying Parallel and Distributed Deep Learning: An In-Depth Concurrency Analysis. *ArXiv:1802.09941 [Cs]*.

Bergstra, J., Breuleux, O., Lamblin, P., Pascanu, R., Delalleau, O., Desjardins, G., Goodfellow, I., Bergeron, A., Bengio, Y., & Kaelbling, P. (2011). Theano: Deep Learning on GPUs with Python. *Journal of Machine Learning Research*, *1*, 1–48.

Chollet, F. (2015). Keras. *Https://Keras.Io*.

Craik, A., He, Y., & Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: A review. *Journal of Neural Engineering*, *16*, 031001.

Dai, J., Li, Y., He, K., & Sun, J. (2016). R-FCN: Object Detection via Region-based Fully Convolutional Networks. *Advances in Neural Information Processing Systems 29*, 379–387.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Deng, L. (2014). A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing*, *3*(e2).

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97.

Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., Bates, S., Bhatia, S., Boden, N., Borchers, A., Boyle, R., Cantin, P., Chao, C., Clark, C., Coriell, J., Daley, M., Dau, M., Dean, J., Gelb, B., Yoon, D. H. (2017). In-Datacenter Performance Analysis of a Tensor Processing Unit. *Proceedings of the 44th Annual International Symposium on Computer Architecture*, 1–12.

Khan, H., Marcuse, L., Fields, M., Swann, K., & Yener, B. (2018). Focal Onset Seizure Prediction Using Convolutional Networks. *IEEE Transactions on Biomedical Engineering*, *65*(9), 2109–2118.

Kitayama, M., Otsubo, H., Parvez, S., Lodha, A., Ying, E., Parvez, B., Ishii, R., Mizuno-Matsumoto, Y., Zoroofi, R. A., & Snead, O. C. (2003). Wavelet analysis for neonatal electroencephalographic seizures. *Pediatric Neurology*, *29*(4), 326–333.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 1097–1105.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, *1*(4), 541–551.

Long, J., Shelhamer, E., Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440.

Ozcan, A. R., & Erturk, S. (2019). Seizure Prediction in Scalp EEG Using 3D Convolutional Neural Networks With an Image-Based Approach. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *27*(11), 2284–2293.

Perslev, M., Jensen, M., Darkner, S., Jennum, P.J. rgen, Igel, C. (2019). U-Time: A Fully Convolutional Network for Time Series Segmentation Applied to Sleep Staging, *Advances in Neural Information Processing Systems,* 4417–4428.

Qu, S., Li, J., Dai, W., & Das, S. (2016). Learning Filter Banks Using Deep Learning For Acoustic Signals. *ArXiv:1611.09526 [Cs]*.

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., & Faubert, J. (2019). Deep learning-based electroencephalography analysis: A systematic review. *Journal of Neural Engineering*, *16*(5), 051001.

Sainath, T. N., Kingsbury, B., Mohamed, A., & Ramabhadran, B. (2013). Learning filter banks within a deep neural network framework. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 297–302.

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv:1409.1556 [Cs]*.

Temko, A., & Lightbody, G. (2016). Detecting Neonatal Seizures With Computer Algorithms. *Journal of Clinical Neurophysiology*, *33*(5), 394–402.

Thodoroff, P., Pineau, J., Lim, A. (2016). Learning Robust Features using Deep Learning for Automatic Seizure Detection, *Machine Learning for Healthcare Conference*, 178–190.

Truong, N. D., Nguyen, A. D., Kuhlmann, L., Bonyadi, M. R., Yang, J., Ippolito, S., & Kavehei, O. (2018). Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. *Neural Networks*, *105*, 104–111.

Ullah, I., Hussain, M., Qazi, E.-H., & Aboalsamh, H. (2018). An automated system for epilepsy detection using EEG brain signals based on deep learning approach. *Expert Systems with Applications*, *107*, 61–71.

Webber, W. R. S., Lesser, R. P., Richardson, R. T., & Wilson, K. (1996). An approach to seizure detection using an artificial neural network (ANN). *Electroencephalography and Clinical Neurophysiology*, *98*(4), 250–272.

Yao, R., Lin, G., Shi, Q., & Ranasinghe, D. (2017). Efficient Dense Labeling of Human Activity Sequences from Wearables using Fully Convolutional Networks. *ArXiv:1702.06212 [Cs]*.

Zhao, D., Chen, Y., & Lv, L. (2017). Deep Reinforcement Learning With Visual Attention for Vehicle Classification. *IEEE Transactions on Cognitive and Developmental Systems*, *9*(4), 356–367.

Zheng, W.-L., & Lu, B.-L. (2015). Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks. *IEEE Transactions on Autonomous Mental Development*, *7*(3), 162–175.

Chapter 5: Architecture optimisation and efficient data usage

5.1 Introduction

In Chapter 4, it was experimentally demonstrated that a neonatal seizure detection algorithm can be developed without the need for hand engineered features to represent EEG windows. A convolutional neural network (CNN) architecture was developed which acted as both feature extractor and classifier. In this chapter, a fully convolutional architecture which learns to detect seizures from temporal EEG signals is designed. The networks trained in this exploration equal and surpass state-of-the-art performance levels.

The architectures developed in this chapter benefit from three major innovative steps. First, the developed system utilises short sample-size filters which are configured in a hierarchical manner to learn rich signal representations through multiple non-linear transformations. Second, the architecture is designed to be fully convolutional so that the weights of convolutional layers (filters) are trained to parameterise the signal properties with a discriminative target in mind. Section 5.3 investigates various architectural choices made throughout the design process and explores how they influence the network performance.

The third innovative step is the incorporation of seizure location invariance into the system. Section 5.4 demonstrates that the designed deep learning architecture can effectively learn from weakly labelled data. Deep learning algorithmic performance is typically considered to be a function of the amount of training data. In contrast to this, the performance of traditional machine learning algorithms is dependent on of the quality of engineered features. The ability to learn from weakly labelled datasets has been previously demonstrated in the image processing and audio processing domains (Kumar and Raj, 2016). The ability to learn from weakly labelled neonatal seizures in multi-channel EEG increases the amount of training data available to the network.

The performance of the developed system is assessed using cross-validation on a large database of continuous EEG recordings of 834 hours in duration (Term Dataset I), is further validated on a held-out publicly available dataset of 94 hours in duration (Term Dataset II), and compared with two baseline SVM based systems (Tapani et al., 2019; Temko et al., 2013).

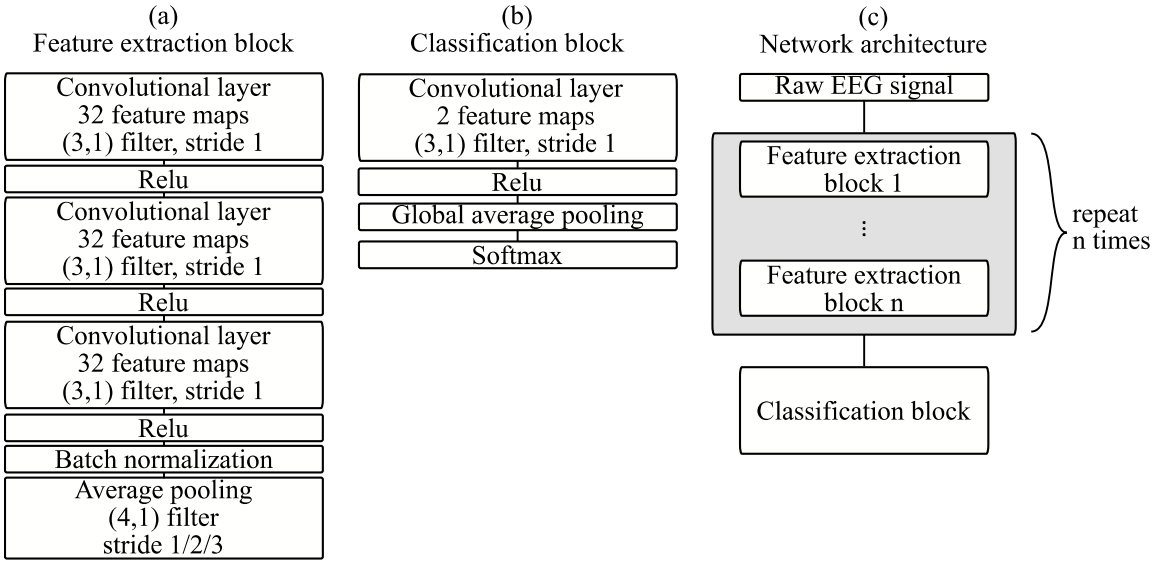


Figure 5.1 The structure of each feature extraction block (a) and classification block (b) used in experiments. The full network architecture is shown in (c). Varying the number of feature extraction blocks and the pool stride value in each block controls the network complexity and the size of receptive field in the final convolutional layer.

5.2 Network architecture and experimental procedure

The experiments in this chapter can be separated into two stages: the first stage utilises a 1-dimensional (1D) fully convolutional network (FCN) architecture to detect seizures in single channel EEG; the second stage utilises a 2-dimensional (2D) FCN architecture to detect seizures in multi-channel EEG. Both experimental procedures employ the same basic architectural building blocks, which are outlined in Section 5.2.1. The train, validation, and test splits, and the training hyperparameters for these experimental procedures are outlined in Section 5.2.2.

5.2.1 Building blocks of the FCN architecture

In order to create a deep learning architecture for neonatal seizure detection, two important architectural blocks need to be defined using deep learning layers. These blocks are referred to as the feature extraction and the classification blocks, Figure 5.1 shows their structures. The names of the blocks are primarily based on their location in the architecture rather than the constituent layers since the only layers required to implement these blocks are convolutional normalisation, activation, and pooling layers. Figure 5.1 (a) is the feature extraction block of the deep CNN architecture. These blocks can be stacked to increase network depth. Figure 5.1 (b) shows the classification block of the deep CNN architecture which allows the learned intermediate representations of the EEG to be classified as either seizure or non-seizure.

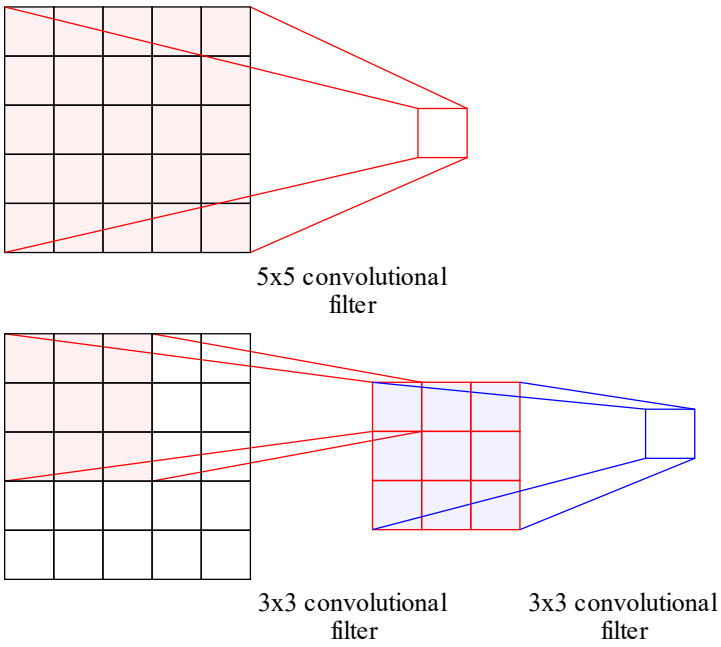


Figure 5.2 Comparison of convolutional filter sizes. A 5x5 convolutional filter operation requires 25 trainable parameters. The two sequential 3x3 convolutional filter operation require 18 trainable parameters. Typically all convolutional operations are followed by a non-linear activation function. The smaller convolutional filters require fewer parameters but potentially have more discriminative power due to the two non-linear operations.

In a typical CNN architecture, the classification stage is implemented using fully connected layers. In an FCN architecture, convolutional layers are used both to extract features from inputs and to perform classification, as illustrated in Figure 5.1. The full network architecture, with a variable number of feature extraction blocks, is shown in Figure 5.1 (c). The number of convolutional filters in the classification block is chosen to be equal to the number of classes. A global average pooling operation is applied to each of the feature maps in the final layer. This is followed by a softmax activation function that converts the values into a set of probabilistic outputs. In this manner, since there are no fully connected layers, the network is optimised to extract the necessary discriminative information through the hierarchy of convolutional filters before reaching the global average pooling operation, thus learning the 'seizure-ness' and 'background-ness' of the input raw EEG in the final convolutional layer.

The convolutional layers in the feature extraction and classification blocks all have small filters which are three samples wide. Small filters have been widely used in deep learning architectures as they facilitate the learning of a wide range of feature resolutions, are more computationally efficient than larger filters, and are considered to be more discriminative (Lee et al., 2018; Simonyan & Zisserman, 2015). Figure 5.2 compares a single 2D 5x5 filtering

operation with two sequential 2D 3x3 filtering operations and demonstrates how this results in more non-linear decision making layers from less trainable parameters over the same area of input data. The parameter reduction property does not translate to 1D convolutional operations. However, in 1D convolutional operations sample sized filters offer more variation between hierarchical receptive fields and increased non-linear decision making layers for the same receptive field input.

FCN architectures offer some benefits when compared to a CNN that relies on fully connected layers. First, the number of network parameters is drastically reduced, which along with weight sharing eliminates the need for regularisation routines such as dropout or L1-L2 regularisations. The absence of a fully connected layer allows the network to classify inputs of any length since the last global average pooling layer can process a sequence of any length. Finally, yet importantly, since only convolutional layers are used, the temporal relationship between the input and the output is preserved until the final global pooling layers. Maintaining ordering in time throughout the classification network can localise the most interesting activity by analysing the output of the last convolutional layer, this is explored further in Chapter 7. In computer vision, FCN architectures have been used for object detection and localisation tasks with state-of-the-art results (Dai et al., 2016; Yao et al., 2017).

5.2.2 Experimental procedure: cross-validation and training hyperparameters

Classifiers need to be validated in a patient independent framework in order to properly assess their ability to perform on unseen data. The leave-one-out (LOO) cross-validation method is a suitable test procedure for a patient independent system. In this method all patients except one are used for training, with the trained classifier then being tested by checking its performance on the one remaining unseen patient. This routine repeats until every patient has been treated as the held-out test patient. All experiments follow the same LOO testing routine. In this work, the area under the receiver-operating curve (AUC) is the primary measure of classifier performance.

The need for strong, channel-specific labels during the training of the 1D FCN architecture limits the amount of Term Dataset I that can be used for training. Each infant has a limited amount of seizure EEG data which have been clearly attributed with confidence to a particular channel; these are strong labels. All strongly labelled data are utilised during training. To balance the classes in the training dataset not all the available non-seizure EEG data are

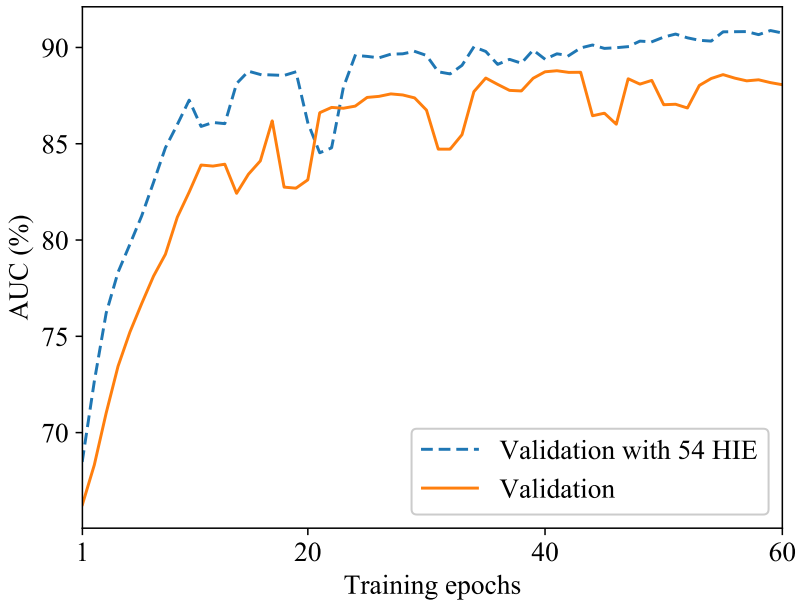


Figure 5.3 Learning curves which demonstrate how validation AUC score is affected by the inclusion of EEG from 54 infants who experienced HIE in the training dataset. The solid orange line plots the validation AUC score for each epoch of training where the training dataset only consists of data from Term Dataset I. The dashed blue line plots the validation AUC score for each epoch of training where the HIE EEG data have been added to the same training dataset.

included in the training dataset. A randomly selected sample of the non-seizure epochs, which is twice the amount of strong seizure labels for that infant, is included in the training dataset.

To increase diversity in the dataset, 54 files, each approximately one hour in duration, from infants who experienced HIE but did not have any recorded seizure events, are also added to the training dataset. The learning curves shown for a validation set in Figure 5.3 emphasise the impact this that increased diversity has on classifier performance. The added HIE data adds no new information about seizure patterns to the training data, but it does increase the variety of background EEG morphology in the training dataset.

Since only a small portion of EEG from each infant is utilised for training the channel specific algorithm (approximately 2% due to the scarcity of strong labels), the remaining EEG from the 17 training infants can be used in the validation set for early stopping, to prevent over-fitting. When the validation score stops improving for more than 25 iterations the training stops. The network architecture that gave the best validation score is then used for testing. The validation performance is measured against the fused probabilistic outputs from all EEG channels; the maximum operation is applied across the channel dimension to fuse channels. The fusion of probabilities from channels in the validation routine means that the weak labels can be utilised

to measure validation performance, therefore all of the data from the 17 training infants are used to measure the performance during validation.

The optimization routine uses stochastic gradient descent with a learning rate of 0.001 and Nesterov momentum set to 0.9. The training batch size is 4096. Each training example is a 256x1 vector, this allows the network to be trained with a relatively large batch size when compared to image processing examples where the inputs are 2D images with an additional colour (red-green-blue) dimension (Goyal et al., 2018). The loss function is binary cross-entropy.

Later in this chapter, in experiments with the 2D FCN architecture, the entire recordings from 17 training infants can be used for training since all of the data has weak labels. In this process (while one infant is used for testing, as per the LOO routine), three infants out of 17 are used for validation and early stopping, the remaining 14 infants are used for network training. To balance the target classes, the seizure and non-seizure training examples are weighted in the loss function during training to simulate a balanced dataset. Since the division of training data to train and validation sets is no longer deterministic, the split is repeated three times by randomly choosing 3 validation infants from 17. For each split, when the validation score stops improving for more than eight iterations the training stops. The set of trained weights from the training iteration that resulted in the highest validation AUC score is utilised in the saved trained network. This routine is repeated three times with a different random validation split each time; this results in three separate models, each based on a slightly different set of training infants. To fully utilise all three models, they are combined using an ensembling technique. The average of the probabilistic outputs from the three trained networks was taken to be the final probabilistic output at the inference stage.

The optimization routine uses stochastic gradient descent with a learning rate of 0.001 and Nesterov momentum set to 0.9. The training batch size is 300. Each training example is a 256x8 array due to the multi-channel EEG representation, the training batch size was reduced to account for the increased input size. In 2D FCN experiments, the binary cross-entropy loss function is also employed.

5.3 1D FCN architecture: exploring the effects of network architecture

The 1D FCN model is designed to receive single channel EEG as input and to apply the convolutional filters across the temporal dimension of the input signal. The model is limited to single channel inputs for two reasons.

First, the initial deep learning model was limited to single channel inputs to facilitate a fair comparison with the baseline SVM algorithm; the baseline SVM algorithm was trained to predict seizures from single channel EEG windows. The Cork SVM classifier requires strong, channel-specific labels during the training stage and the developed algorithm is applied to each EEG channel separately; so, to compare classifier performances this aspect of the algorithm should remain the same.

Second, single channel inputs are utilised to reduce the label noise and to take advantage of the available channel specific seizure annotations. It is known that seizures in neonatal EEG can be focal in nature, often appearing in only a single channel or a subset of channels. Term Dataset I has weak labels for all files for all seizure events, but this dataset only has strong channel specific labels for a small subset of seizure events. Furthermore, for the seizure events which have strong labels, not all channels are labelled so it is not known which channels represent non-seizure during a seizure event. The lack of strong, channel-specific labels motivates the development of a 1D neonatal seizure detection FCN.

To create an architecture which processes one EEG channel at a time all the convolutional and pooling filters must be 1D. In this work, network depth can then be arbitrarily chosen by varying the number of feature extraction blocks, see Figure 5.1 (c). More blocks in the network allows for the learning of deeper networks with increased feature complexity. The pool stride in each feature extraction block refers to the sample stride between successive pooling operations within the feature extraction blocks. Increasing the stride has a down-sampling effect, reducing the number of samples in subsequent layers.

A convolutional filter in each convolutional layer is influenced by a certain span of EEG samples in the input layer. The receptive field (RF) is the range of samples in the input signal corresponding to a convolutional operation in a particular layer. In the first convolutional layer, the RF of each convolutional operation is equal to the size of the convolutional weight matrix, but in deeper layers, the RF widens. The range of samples in the input layer that is seen by a filter can be increased by adding convolutional layers or by down-sampling using pooling layers. The maximum possible RF in the last layer in these experiments is 256 samples of EEG (8-seconds at a sampling rate of 32Hz). The 8-second window size with 1-second shift was chosen based on prior work, and it is maintained at this value to ensure the comparability.

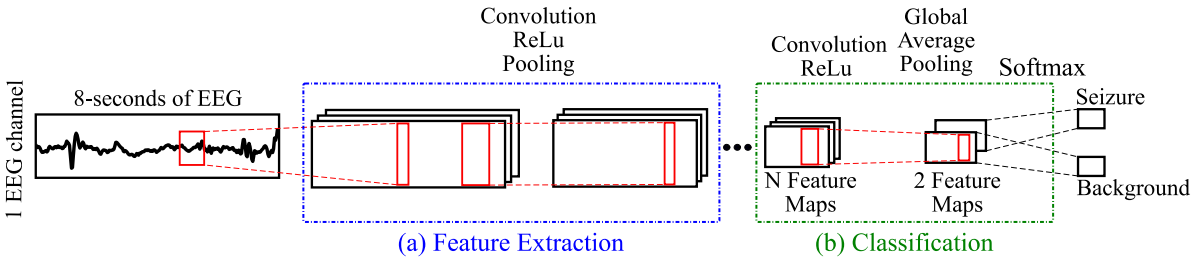


Figure 5.4 The 1D FCN designed for the seizure detection task. This network relies only on convolutional, pooling and batch normalisation layers. The convolutional layers make up the feature extraction section (a) of the network; they can have any amount of feature maps; in this work the number of feature maps is set to 32, unless otherwise stated. The final convolutional layer, which learns to classify the inputs into seizure or background categories, has the same number of feature maps as network output categories (b). After a global pooling layer, these final feature maps connect directly to the individual network outputs. (O'Shea et al., 2020)

The RF in each layer can be computed recursively as:

$$RF_l = (RF_{l-1} - 1) \times s + k, \tag{5.1}$$

where layer l-1 precedes layer l, s is the stride and k is the filter width of either a convolutional or pooling layer. In this work, the pool stride takes values of one, two or three samples. From Equation 5.1, the filter stride has a multiplicative influence on the RF width, while the filter width has an additive influence.

In CNN architectures employing fully connected layers, the value of the receptive field in the final convolutional layers is not critical; the presence of dense connections is the equivalent of having a receptive field that spans the entire input. In the FCN architecture proposed in this chapter, the final convolutional layer has the widest receptive field in the network. The final layer convolutional operations are the most complex as they span the greatest number of input samples. Thus, the final convolutional layer acts as the classification layer and its RF length determines how much of the input EEG is seen by each element before global average pooling.

The previous subsections describe the building blocks needed to perform feature extraction and classification. The first model in this work utilises a 1D FCN architecture which is presented in Figure 5.4. Figure 5.4 also illustrates how the final convolutional layers in the FCN are designed to act as classification layers. The EEG input is represented as a 1D temporal waveform. The input is an 8-second window of EEG, sampled at 32Hz, resulting in a 256x1 input vector. Figure 5.5 (right path) positions the designed 1D FCN architecture with respect to the SVM-based system (Temko et al., 2013) – similarly, the 1D FCN operates on a single

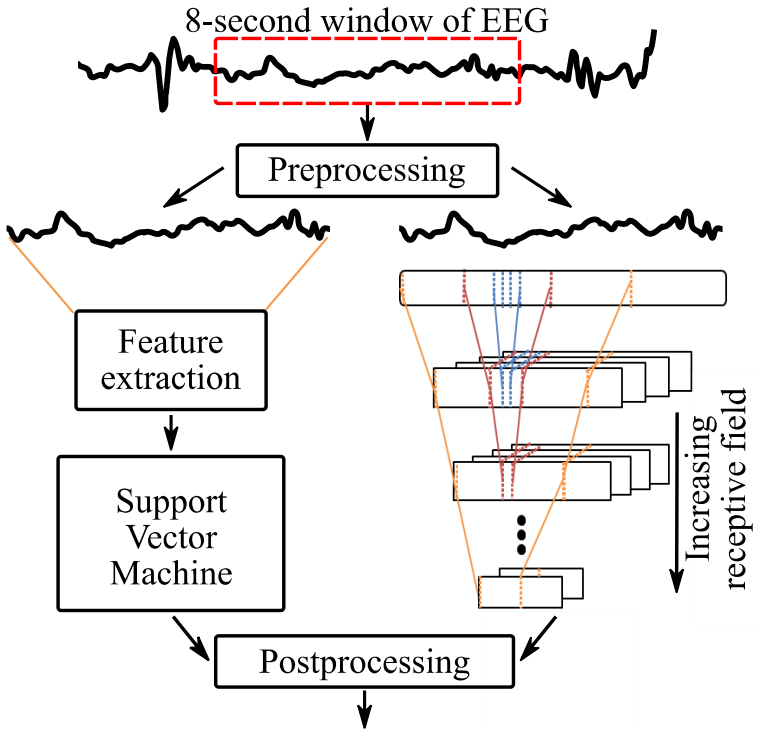


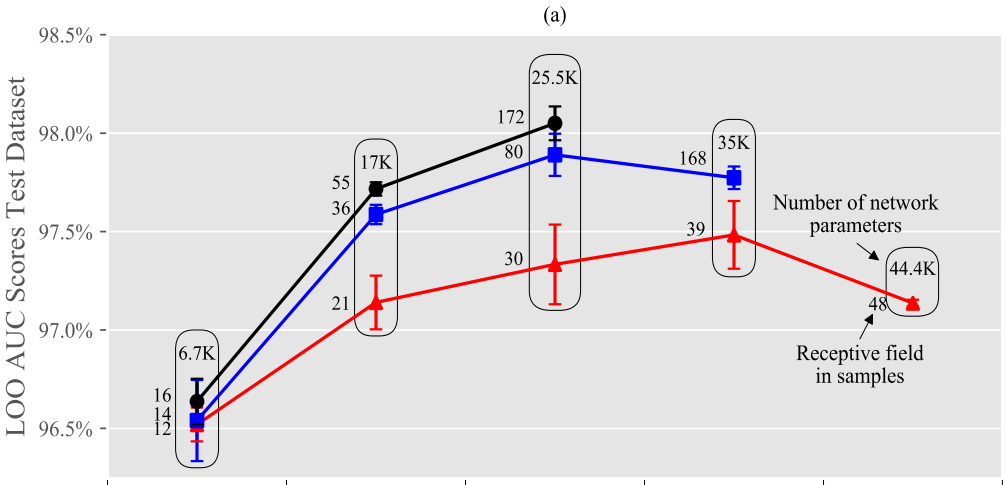
Figure 5.5 A comparison of the SVM Cork algorithm (left path) and the deep learning architecture (right path) developed in this work (Temko et al., 2013). Both algorithms process an 8-second window of EEG. The same postprocessing and preprocessing routines are applied. The 1D FCN directly substitutes for the feature extraction and SVM blocks in the Cork SVM algorithm.

EEG channel and shares the same preprocessing and postprocessing steps but it combines both the feature extraction and classification blocks into one end-to-end optimisation routine. Similar to previous works, the 1D FCN architecture requires strong labels for training (O'Shea et al., 2017; Tapani et al., 2019; Temko et al., 2011, 2013). The use of strong labels implies that only a small sub-set of the data can be used for training.

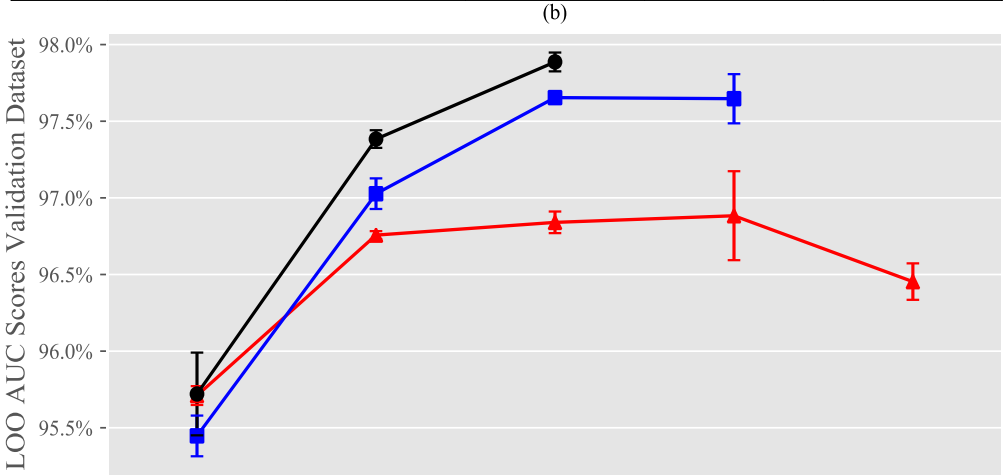
5.3.1 Performance variation: receptive field versus network depth

Figure 5.6 shows how test (a) and validation (b) performances in the LOO experiments vary with different parameter combinations. The two variables in these experiments are the number of feature extraction blocks used, the feature extraction blocks are defined in Figure 5.1 (a), and the stride between the pooling operations in the feature extraction blocks. The horizontal axes represent the experiments with the number of feature extraction blocks gradually increased from one to five. The shallowest architecture exploited in this work uses one convolutional block followed by one classification block (four convolutional layers in total). The deepest architecture used in these experiments has five feature extraction blocks and one classification block (16 convolutional layers in total). The pool stride is incremented from one

to three. Each value of pool stride is represented in Figure 5.6 using a different marker. The vertical axis represents the performance measured as AUC scores.



	1	2	3	4	5
▲ Pool 1	96.52	97.14	97.33	97.48	97.14
■ Pool 2	96.54	97.59	97.89	97.77	-
● Pool 3	96.64	97.72	98.05	-	-



	1	2	3	4	5
▲ Pool 1	95.71	96.76	96.84	96.88	96.45
■ Pool 2	95.45	97.03	97.65	97.65	-
● Pool 3	95.72	97.38	97.89	-	-

Figure 5.6 The results of experiments comparing various combinations of 1D FCN architectural parameters. The horizontal axis represents the number of feature extraction blocks. The vertical axis represents AUC performance. The three different marker styles each represent a different value of pool stride. In (a) every point has a label that shows the maximum receptive field value for that network architecture configuration. Networks with the same number of trainable parameters are grouped together. (a) shows the performance on the test dataset at the end of the LOO experiment. (b) shows the corresponding performance on the validation dataset. For speed-up purposes validation results (b) exclude time-consuming postprocessing steps whereas test results (a) include postprocessing. For this reason, higher values of performance are observed on test data (a) as compared to validation data (b). (O'Shea et al., 2020)

All experiments were run three times starting from a different random initialisation point. Each point is the mean of three experiments, the error bars are the standard deviations across the three experiments. The combination of the number of feature extraction blocks and the pool stride results in different receptive field widths and different numbers of total parameters for each architecture. In Figure 5.6 (a), each data point is labelled with the value of the receptive field in the final convolutional layer for that architecture. The receptive field increases with added convolutional depth and with the increased pool stride. The number of network parameters is not affected by the pool stride, so points with the same number of feature extraction blocks are grouped by number of parameters. Figure 5.6 (a) is the plot of mean AUC scores on the held-out test dataset. Figure 5.6 (b) shows the corresponding AUC scores on the validation data for each of the experiments in Figure 5.6 (a). The test AUC in Figure 5.6 (a) is calculated after the full postprocessing routine is applied to the probabilistic output.

Figure 5.6 (a) shows that increasing the number of network parameters does not lead to a guaranteed performance improvement. Similarly, the largest number of building blocks does not result in the highest score. At very low receptive field values, too much network complexity has resulted in a decrease in classifier performance, this was observed for experiments with a pool stride of one and number of feature extraction blocks set to five. This specific experimental setup had the largest number of network parameters, but it produced a low performance compared to the other depths investigated.

Most of the improvement in AUC performance in this work is achieved by using wider receptive fields. The receptive field can be increased by adding more feature extraction building blocks or by using a down-sampling operation (pooling). Figure 5.6 (a) clearly shows the importance of the receptive field width on classifier performance.

Originally, the FCN was developed for the task of semantic segmentation of images (Long et al., 2015), and in this image processing domain the effects of receptive field width have been studied more extensively. The inclusion of dilated convolutions has also contributed to success in semantic segmentation tasks (Yu & Koltun, 2016). Dilated convolutions give an increase in receptive field without losing resolution or coverage, two very important considerations when trying to densely predict labels. Furthermore, the inclusion of one-sided dilated convolutions has been shown to outperform traditionally used recurrent neural networks in a series of sequence modelling tasks (Bai et al., 2018). Experimental results indicated that the influence of the receptive field in convolutional architectures is more valuable than the architectural

choice of networks specifically designed for sequence modelling tasks. In this work, the importance of the receptive field for a fully convolutional temporal signal classification algorithm was explored; it is the first time this architecture has been proposed for use with physiological signals.

It can be seen from Figure 5.6 (b) that the trends seen in the validation results are reflected in the test results. This implies that the choice of architecture, the depth, the pool stride, etc., can be reliably selected using the validation dataset. When designing algorithms through a LOO process it is fundamental that the architecture selection and tuning routines are not informed by the performance of the algorithm on the held-out test patient but rather driven by the performance on the validation data. To avoid leakage between the training and testing patients a validation set is used to select the best hyper-parameters and dictate when to implement early stopping.

It is also important to highlight that the results in this section are not an exhaustive search of all possible architectures for the 1D FCN. In Figure 5.6, a larger receptive field could be obtained by creating a more flexible building block or combining building blocks with different pool strides. The results presented here aim to show the importance of understanding, when designing a CNN architecture for a new data domain, which architectural features are important, i.e. the receptive field, and which architectural features have less influence on performance, i.e. the depth.

5.3.2 Comparison with the Cork SVM baseline on Term Dataset I

The development of this deep learning model represents a novel step in the development of neonatal seizure detection algorithms, but it needs to be compared with the current state-of-the-art Cork SVM algorithm in order to test whether deep learning gives an improvement over the traditional machine learning approach (Temko et al., 2011).

The Cork SVM algorithm is based on an SVM classifier and was evaluated on Term Dataset I (Chapter 2). In the preprocessing stage, the raw EEG sampled at 256Hz is band-pass filtered between 0.5Hz and 12.8Hz, down-sampled to 32Hz and split into 8-second windows with a 4-second shift. The SVM algorithm relies on 55 features (Chapter 3) extracted from a moving 8-second window of EEG that was selected and engineered for the seizure detection task. A full list of the feature set is reported in Chapter 3. A Gaussian kernel SVM was trained on the extracted features. This training framework requires per-channel annotations (strong labels) in order to create a representation of the seizure data. In the testing stage, which is shown in Figure

5.5 (left path), the trained SVM model is applied to each EEG channel separately with the final decision made after postprocessing and fusing the probabilities across all EEG channels.

The postprocessing stage includes taking a maximum probability across channels, a moving average filter of 60-seconds length, adaptation to the background level of probability and a collar to compensate for the effect of smoothing. One key takeaway from the Cork SVM algorithm is that when many relatively simple EEG signal features are combined with a powerful and robust classifier, such as the SVM, a reliable detector can be trained.

Figure 5.7 shows a comparison between the SVM algorithm and the 1D FCN deep learning algorithm for the LOO experiments for each of the 18 infants in Term Dataset I. The SVM algorithm results presented were reported in prior literature (Temko et al., 2013). The 1D FCN algorithm has three feature extraction blocks and pool stride of three, this architecture was selected using the experimental results in Section 5.3.1. In order to ensure a direct performance comparison, the same preprocessing filters, 8-second windowing and postprocessing steps are reutilised in the FCN algorithm. Both algorithms were trained using the subset of seizures with channel-specific annotations in Term Dataset I. In almost every infant the AUC score was

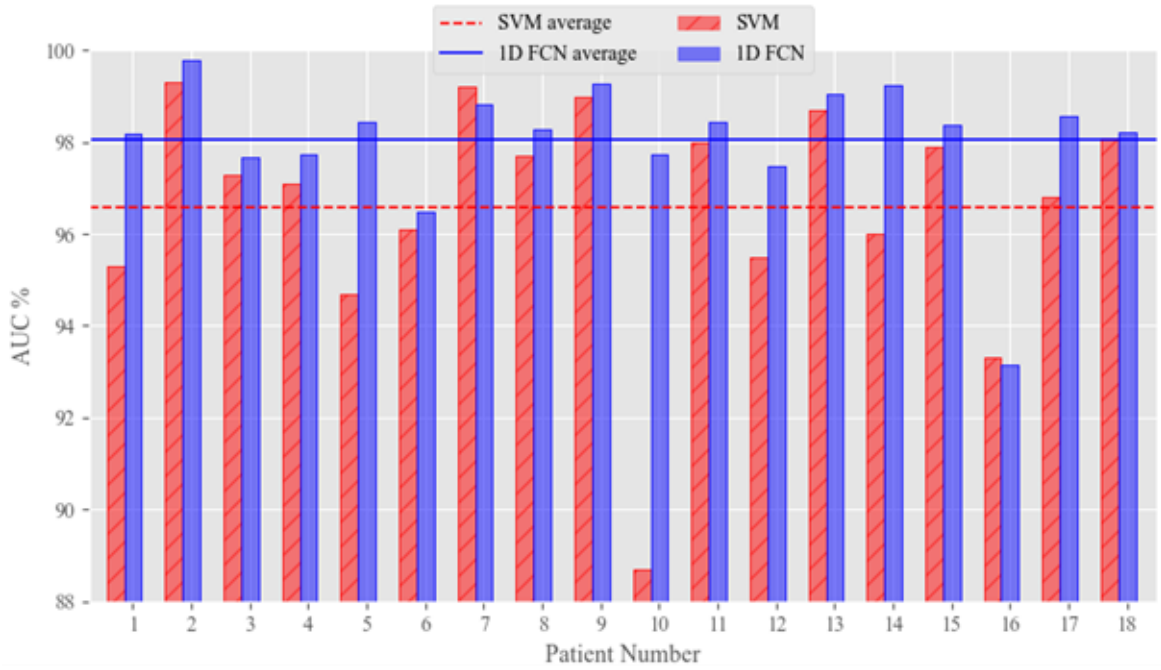


Figure 5.7 A comparison of LOO results for each of the 18 patients in Term Dataset I. Results of the previous state-of-the-art for this dataset, an SVM based architecture which uses 55 features extracted from EEG, are reported alongside the results of the developed deep learning model that was trained using the same set of per-channel annotations. Horizontal lines show the average performance across the 18 infants for the CNN and SVM algorithms. (O'Shea et al., 2020)

improved using the deep learning approach, the average across all AUC scores is improved by over 1%, absolute when using the 1D FCN algorithm.

The direct comparison between the Cork SVM based system and the 1D FCN based system can be performed as both systems use the same strongly labelled training data. Both systems operate on a single EEG channel at a time, share the same preprocessing and postprocessing routines, and are tested in the same LOO fashion on the same dataset. These results have proven the merits of exploring deep learning for the neonatal seizure detection task and have driven this work, which is an exploration of how far deep learning can improve the performance.

The fact that a deep learning architecture has outperformed the SVM algorithm in Figure 5.7 shows that the convolutional layers can extract task appropriate features from the raw EEG signal without the need for hand-crafted features. While the achievements of deep learning algorithms in other signal processing fields certainly created the premise for this study, the ability for the deep learning algorithm to achieve a better AUC score than the highly engineered and tuned SVM algorithm required a rigorous process of architectural design. However, the magnitude of the increase achieved is noteworthy, considering the years of engineering and medical expertise that went into the development and selection of the features in the SVM algorithm. The SVM algorithm is optimised in two stages, first the 55 features are selected to be relevant for the seizure detection task, and next the classification algorithm is trained and optimised to distinguish between seizure and background EEG. Each of these stages required a vast amount of knowledge and time from clinical, engineering, and machine learning experts.

Neonatal EEG is a highly complex signal and it varies widely between patients. Through the extraction of a feature set, some of the signal information contained within the EEG waveform may be lost. The deep learning approach in this work uses the temporal EEG signal; all the information contained in the EEG is used to train the neural network. The fact that the FCNs have outperformed the state-of-the-art algorithm on this large clinically representative dataset proves that the deep learning algorithm was able to extract more meaningful representations from raw EEG than can be done manually when restricted by prior knowledge and expert assumptions such as signal stationarity (e.g. the application of the discrete Fourier transform). The engineer or the clinical expert is using their experience to decide which parts of the data will give the most information when choosing a feature set. In contrast, the FCN has an entirely data driven and objective-driven feature selection routine, as it is part of the backpropagation process. A data driven latent representation is richer than the one that can be derived with

human influence and intervention. An additional advantage is that both the feature extraction process and the input classification process are optimised in one end-to-end routine.

5.3.3 Performance variation: impact of feature maps and network capacity

Previous results have shown that the designed deep learning architecture is on par with a state-of-the-art machine learning algorithm on Term Dataset I. The effect of receptive field size and the number of convolutional layers in the chosen FCN architecture has also been observed. Further experiments were also conducted to investigate the influence of the number of network

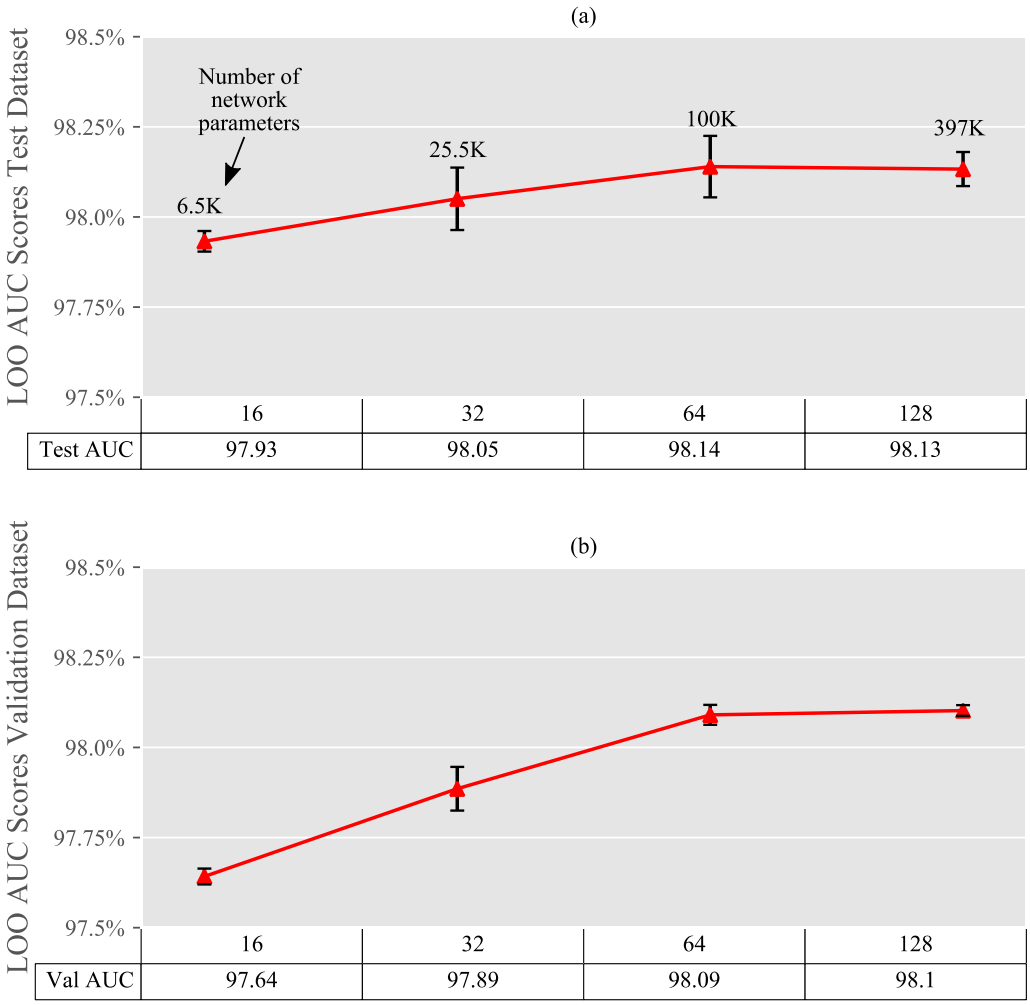


Figure 5.8 The results of experiments where the number of feature maps in each layer of the FCN is varied from 16 to 128. The number of convolutional layers and the pool stride utilised in these experiments was selected based on the results in Figure 5.6. The horizontal axis represents different number of feature maps in each architecture; every convolutional layer has the same number of feature maps. The vertical axis represents AUC performance. The performance on the test set is shown in (a), the corresponding performance on the validation set is shown in (b). In both (a) and (b) the error bars represent the standard deviation of performance across three experiments. The number of network parameters in each architecture is shown above the markers in (a).

parameters on performance by experimenting with different numbers of feature maps in the convolutional layers.

Figure 5.8 shows the results of experiments where the number of feature maps in each convolutional layer is adjusted, and the effect of this change on performance is evaluated. The number of feature maps in each layer is incremented in increasing powers of two (2^4, 2^5, 2^6, 2^7). Figure 5.8 (a) reports the total number of network parameters for each network architecture with varying numbers of feature maps above the markers.

The performance of the FCN network increases with increasing number of network parameters. The network has more weights and biases which can be adjusted during the training routine to differentiate between seizure and non-seizure epochs. The network utilised for testing in Figure 5.8 is the 1D FCN with three feature extraction blocks and a pool stride of three, i.e. the same network architecture that was selected through experiments shown in Figure 5.6. These experiments show that the selected architecture can be further optimized for the neonatal seizure detection task by experimenting with different model hyper-parameters.

The variation in validation AUC scores in Figure 5.8 (b) shows how the number of trainable parameters in each convolutional layer influences the performance of the trained FCN network. Increasing the number of feature maps in each layer results in more accurate classifiers, but the performance increase saturates at higher numbers of network parameters. Too many trainable weights in the FCN architecture could result in network overfitting, where an over-parameterised architecture no longer generalises to unseen data. This may explain the slight decrease in performance seen in the test AUC for the 128 feature map architecture in Figure 5.8 (a).

The increase in parameters from 6,500 (16 feature maps) to 25,500 (32 feature maps) results in an 0.12% improvement in test AUC performance. The increase in parameters from 25,500 (32 feature maps) to 100,000 (64 feature maps) results in a smaller 0.09% increase in test AUC performance. Knowledge of this diminishing returns behaviour should inform the development of neonatal seizure detection algorithms that are memory efficient, computationally efficient, and power efficient. Further experiments in this chapter will utilise the network which has 32 feature maps in each layer.

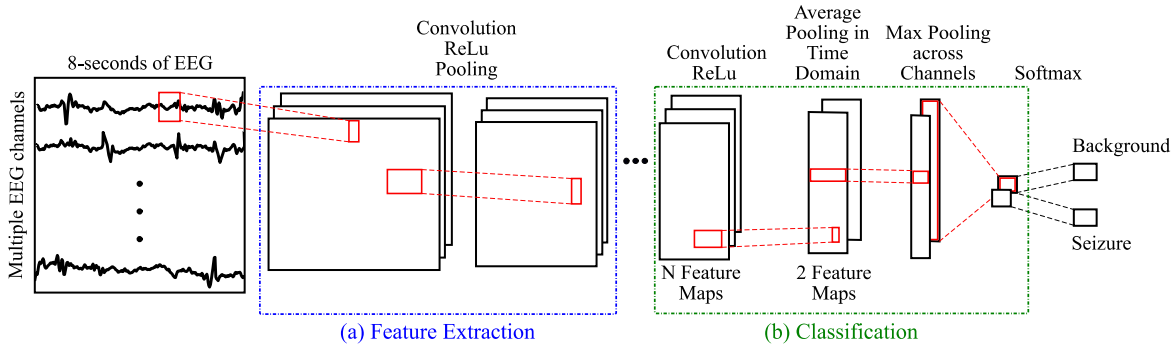


Figure 5.9 The 2D FCN designed for the seizure detection task. This network is designed to take multiple channels of EEG as input and it outputs the seizure decision across all channels. An average pooling layer is designed to create a value which corresponds to the level of seizure-ness in each channel and a maximum pooling layer recognises if there is a seizure in any one channel. (O'Shea et al., 2020)

5.4 2D FCN architecture: levering more data by using weak labels

An alternative deep learning architecture is also explored in this work, which is shown in Figure 5.9. This architecture is a 2D FCN that takes multiple EEG channels as input and detects if a seizure happens in any one of the channels. The input can take an arbitrary number of EEG channels; this number does not have to be consistent across training and testing. In this work, each input channel is an 8-second window of EEG sampled at 32Hz, resulting in a 256xN array, where N is the number of input channels. The main architectural difference with respect to the 1D FCN architecture is the inclusion of a final maximum pooling layer, which takes the maximum across feature maps from all EEG channels. This max pooling layer allows one of the postprocessing steps (channel fusion) required in the 1D FCN to be incorporated into the end-to-end learning routine.

The 2D FCN architecture is designed using the same layers as the 1D FCN network with all filters being 1D, which implies that no inter-channel correlations are considered to avoid learning patient specific seizure location dependencies. Instead, this architecture is designed to learn the presence of seizure patterns among all available channels, regardless of the seizure location. The 2D FCN network now must accomplish a much more complex task – to learn to distinguish between seizure activity and non-seizure activity from weakly labelled data. Weak labels in Term Dataset I may have up to 1-to-7 signal-to-noise ratio – e.g. when a seizure is only present in a single channel out of 8 EEG channels. All 8 channels are presented to the network and the same convolutional filter will be tuned with EEG samples from all signals. The main advantage of this architecture is that it does not require the seizure labels to be channel-specific; this means that the entire Term Dataset I could be used for training instead

Table 5.1 Mean and standard deviation of AUC scores across the 18 infants in Term Dataset I. The Cork SVM which was previously developed is compared with the two classifiers developed in this work (Temko et al., 2013). (O'Shea et al., 2020)

		SVM	1D FCN	2D FCN
AUC (%)	mean	96.6	98.1	98.5
	std	2.5	1.4	1.1
AUC90 (%)	mean	82.9	86.9	88.2
	std	8.4	7.5	8.1

of the subset that has channel-specific labels. The hypothesis is whether the availability of over 40 times more training data (in terms of the number of EEG samples) can compensate for using noisier weak labels.

Table 5.1 shows the mean AUC and AUC90 LOO scores for the SVM Cork algorithm, the 1D FCN and the 2D FCN algorithm that correspond to the best performing architecture in Figure 5.6 (selected on the validation data), this is also the architecture with 32 feature maps in Figure 5.8. The SVM and 1D FCN are trained using the same subset of Term Dataset I which has channel-specific annotations and the 2D FCN is trained using the entirety of Term Dataset I (available for training after the LOO split). Table 5.1 also shows the standard deviations of the scores across the 18 infants in each LOO experiment. The relative performance improvement achieved using deep learning with respect to the Cork SVM baseline is $\frac{98.5-96.6}{100-96.6}\%=56\%$.

5.4.1 Comparison with the Helsinki SVM baseline on Term Dataset II

This section will explore performance of the deep learning algorithms on the publicly available Term Dataset II (Chapter 2) and a comparison with the publicly available Helsinki SVM algorithm (Chapter 3). The validation of an algorithm's performance on an unseen dataset is a good indicator of its generalisation ability. Table 5.2 illustrates the performances of the 2D FCN developed in this work and the SVM Helsinki (Tapani et al., 2019), allowing the generalisation abilities of both algorithms to be compared in matched and mismatched conditions. The matched condition is when the source of the dataset used for training and testing of the algorithm is the same. This implies that the properties of the dataset do not change from train time to test time – the variables such as the seizure burden, inter-observer variability in annotations, overall duration of recordings, montage setup, electrode application procedure, recording equipment characteristics with built-in proprietary filters, etc. are kept constant. The mismatched condition occurs when these properties change between train time and test time.

Table 5.2 Comparison of LOO performance and held out test set performance for the 2D FCN algorithm trained on Term Dataset I, which is the dataset used in this work, and the Helsinki SVM algorithm trained on Term Dataset II. The AUC-concatenated results of training on Term Dataset I or Term Dataset II data and then testing on each dataset are shown. The test type, either Leave One patient Out (LOO) or a once-off unseen test (Test), is reported in brackets alongside each score. (O'Shea et al., 2020)

AUC (%)		
	Test Database	
Algorithm	Term Dataset I	Term Dataset II *
Helsinki SVM	(Test) 89.4	(LOO) 95.5**
2D FCN	(LOO) 98.5	(Test) 95.6

* AUC is reported on concatenated data so that all infants can be included in the calculation
** As reported in (Tapani et al., 2019)

The 2D FCN has the best performance in both conditions. In addition, the drop in performance between matched and mismatched conditions is smallest for 2D FCN – it achieves a much higher AUC performance on Term Dataset II than the Helsinki SVM algorithm achieves on Term Dataset I. Overall it shows that the data-driven optimisation of feature extraction and classification is not only robust to variabilities but also results in a more accurate classifier which has the ability to generalise across different data sources, and different seizure and background characteristics.

Term Dataset II contains data from a large cohort of infants, with EEG from 79 different infants. However, each infant is represented by short duration EEG segments; on average 85 minutes. These short windows of EEG are not representative of the EEG recordings typically required in the NICU. Another important note is that the median postnatal age of the infants at the time of recording was three days. In infants who have experienced a hypoxic ischaemic event, most seizures begin in the first 24 hours after birth, and by day three the seizure burden has dropped by over 90% (Abend et al., 2018). In Term Dataset I, used to train the FCNs in this work, the average postnatal age at recording start is less than 17 hours. The FCN algorithm was trained on long recordings of EEG with relatively low seizure activity; these conditions are typically present in the NICU. The postprocessing which is included in the FCN algorithms was optimised with the typical NICU conditions in mind. The postprocessing, which is described in more detail in Chapter 3, includes a moving average filter of 60 seconds and long-term background adaptation (Temko et al., 2013).

The SVM Helsinki was trained on shorter EEG segments that had a relatively high seizure burden. The postprocessing in this algorithm, which applies a decision window of 48-seconds, was tuned to deal with shorter recordings that have a higher proportion of seizure epochs i.e. the conditions of Term Dataset II. Both algorithms achieve an acceptable AUC value in their once-off unseen dataset tests, considering they were trained to deal with very different recording environments. The clinical differences between the dataset used in this work for training and for testing explain the variation between the algorithmic performances when tested on an unseen dataset of different clinical characteristics. The publication and release of a publicly available database containing annotated neonatal seizures is an important step forward in this research field. However, a dataset with longer recordings would be more reflective of the situation that a seizure detection algorithm would be faced with in a NICU.

It should be noted that there is a large variation between the processing times required to implement the inference stage of each algorithm. The Helsinki SVM algorithm relies on the extraction of computationally intensive time-frequency features. In experiments, on the same PC and using the CPU for all computation, the Helsinki SVM takes over 4.5 hours to process one hour of EEG and the 1D FCN and the 2D FCN both take less than one minute to do the same task. When considering the clinical implementation of a neonatal seizure detection algorithm, the processing needs to be realisable in real-time. In Chapter 7, the requirements and trade-offs associated with implementing these fully convolutional architectures on mobile devices to provide real-time decision support are explored. Previous work in this area, has proven that fully convolutional architectures are suitable for low-power, mobile device implementation (O'Sullivan et al., 2018).

5.5 Conclusion

This chapter has presented a framework for developing a neonatal seizure detection algorithm through end-to-end optimisation of the feature extraction and classification stages. The chosen deep learning algorithm is a fully convolutional neural network which operates on raw signals; the use of this architecture is novel in the area of neonatal physiological signal processing.

First, it has been shown that varying the architectural parameters of the FCN that control the complexity and size of the internal learning representation within the network and the width of the receptive field in the final convolutional filters, has the biggest influence on network performance when compared to the network depth. Second, it has been shown that given the same initial conditions the 1D FCN algorithm, which learns a representation from large

amounts of raw, annotated data, outperforms an SVM-based system, which uses a set of hand-crafted features. Third, it has been shown that the designed novel 2D FCN architecture has achieved superior performance by exploiting much larger amounts of data with weak labels. By designing an algorithm which can learn from weakly labelled data this work hopes to accelerate the development of algorithms in the area of neonatal EEG by circumventing the need for laborious, precise, per-channel labelling of datasets. Finally, experiments have shown that the performance can be further improved by increasing the capacity of the network through training more filters in each convolutional layer.

The results in this chapter have shown that the developed deep learning algorithms can outperform two highly-engineered machine learning based algorithms, which were considered to be the state-of-the-art in neonatal seizure detection. Previous algorithms are utilised as comparative baselines in this chapter. The inter-observer agreement is an important aspect to consider when working in the area of neonatal seizure detection (Stevenson et al., 2015). No two neonatal electro-encephalographers will annotate long-term EEG recordings with 100% agreement. For an automated seizure detection algorithm to be clinically useful its level of performance should match that of a human expert annotator i.e. it should have a comparable inter-observer agreement when compared with human experts which has been reported as 83.0% for seizures and 99.7% for non-seizure (Stevenson et al., 2015).

The experiments in this chapter, which explore how increasing the number of feature maps in each layer affects the FCN performance indicate that the networks can be further optimized through hyper-parameter exploration. In machine learning, it is often assumed that increasing the capacity of an algorithm will result in better performance until the network starts to overfit. Bearing this in mind the results in Figure 5.8 are expected. But the results in Figure 5.6, which show that through increasing pool stride, with no increase in network capacity or computational load, the performance of the algorithm can be significantly improved are not intuitive and may be unexpected for some readers.

The results of the 2D FCN algorithm on the publicly available Term Dataset II has indicated its superior performance to an existing published alternative. The performance reported in this work shows that an algorithm based only on deep learning is capable of outperforming algorithms that rely on complex feature sets without the computational load typically associated with intensive feature calculations. The fact that the developed FCN has a

comparatively low number of parameters and computational load makes it suitable for low-memory, low-power applications.

The utilization of 2D raw temporal EEG as FCN input and the associated utilization of weak seizure labels for algorithm training in this chapter is an important breakthrough in the domain of neonatal seizure detection. The process of creating channel-specific, strong seizure labels is a large burden for neurophysiologists and the laborious nature of this process has limited the development of neonatal seizure detection algorithms. The ability to learn from weakly labelled data means that more datasets can be leveraged in the future for neonatal seizure detection algorithm development. The results in this chapter would be applicable to other domains where strongly labelled datasets of temporal signals are not readily available.

5.6 References

Abend, N. S., Jensen, F. E., Inder, T. E., & Volpe, J. J. (2018). Chapter 12—Neonatal Seizures. In J. J. Volpe, T. E. Inder, B. T. Darras, L. S. de Vries, A. J. du Plessis, J. J. Neil, & J. M. Perlman (Eds.), *Volpe's Neurology of the Newborn (Sixth Edition)* (pp. 275-321.e14). Elsevier.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *ArXiv:1803.01271 [Cs]*.

Dai, J., Li, Y., He, K., & Sun, J. (2016). R-FCN: Object Detection via Region-based Fully Convolutional Networks. *Advances in Neural Information Processing Systems 29*, 379–387.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., & He, K. (2018). Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *ArXiv:1706.02677 [Cs]*.

Kumar, A., & Raj, B. (2016). Audio Event Detection using Weakly Labeled Data. *24th ACM International Conference on Multimedia*, 1038–1047.

Lee, J., Park, J., Kim, K. L., & Nam, J. (2018). SampleCNN: End-to-End Deep Convolutional Neural Networks Using Very Small Filters for Music Classification. *Applied Sciences*, *8*(1), 150.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440.

O'Shea, A., Lightbody, G., Boylan, G., & Temko, A. (2020). Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture. *Neural Networks*, *123*, 12–25.

O'Shea, A., Lightbody, G., Boylan, G., & Temko, A. (2017). Neonatal seizure detection using convolutional neural networks. *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6.

O'Sullivan, M., Gomez, S., O'Shea, A., Salgado, E., Huillca, K., Mathieson, S., Boylan, G., Popovici, E., & Temko, A. (2018). Neonatal EEG Interpretation and Decision Support Framework for Mobile Platforms. *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 4881–4884.

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv:1409.1556 [Cs]*.

Stevenson, N. J., Clancy, R. R., Vanhatalo, S., Rosén, I., Rennie, J. M., & Boylan, G. B. (2015). Interobserver agreement for neonatal seizure detection using multichannel EEG. *Annals of Clinical and Translational Neurology*, *2*(11), 1002–1011.

Tapani, K. T., Vanhatalo, S., & Stevenson, N. J. (2019). Time-Varying EEG Correlations Improve Automated Neonatal Seizure Detection. *International Journal of Neural Systems*, *29*(04), 1850030.

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

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

Yao, R., Lin, G., Shi, Q., & Ranasinghe, D. (2017). Efficient Dense Labeling of Human Activity Sequences from Wearables using Fully Convolutional Networks. *ArXiv:1702.06212 [Cs]*.

Yu, F., & Koltun, V. (2016). Multi-Scale Context Aggregation by Dilated Convolutions. *ArXiv:1511.07122 [Cs]*.

Chapter 6: Transfer learning for preterm seizure detection

6.1 Introduction

Seizures are the primary biomarker of neurological dysfunction in term and preterm infants with an incidence of 2-3 per 1000 live births in term infants, and much higher incidences, ranging from 5-30 per 1000 live births, reported in preterm infants (Buraniqi et al., 2017; Lloyd et al., 2017; Shellhaas, 2019). It is critical to detect and treat seizures at the earliest opportunity. The clinical diagnosis of seizures in full term infants is challenging without brain monitoring (Murray et al., 2008; Okumura et al., 2008**).** In preterms, this challenge is much greater due to the prevalence of electrographic-only seizures which show no visible clinical manifestations (Okumura et al., 2008). Identification of seizures with clinical manifestations is complicated by the vast repertoire of jerky movements that are often seen in preterms, many of which are essential for normal sensorimotor development (Khazipov et al., 2004; Milh et al., 2007). These jerky movements can be difficult to distinguish from seizures which can result in unnecessary treatment of infants with anti-epileptic drugs.

Continuous EEG monitoring is the optimal method available for the detection of seizures in infants. Interpretation of neonatal EEG requires highly trained healthcare professionals and as a result, it is limited to specialized centres. Several computer methods have been developed for the detection of seizures in term infants (Ahmed et al., 2017; Ansari et al., 2016; Bogaarts et al., 2016; Deburchgraeve et al., 2008; Gotman et al., 1997; Navakatikyan et al., 2006; Ocak, 2009; O'Shea et al., 2017; Temko et al., 2011; Thomas et al., 2011). Over the past 20 years term EEG seizure detection algorithms (T-SDAs) have passed several stages of development (Temko & Lightbody, 2016). The early algorithms relied on signal processing routines to extract information from the EEG followed by hand-tuned rules and thresholds for decision making. Over many years of research simple features have evolved into more sophisticated hand-crafted characteristics while the rules and thresholds have transformed into data-driven classifier-based decision making. The state-of-the-art results have been reported by a recent deep learning solution which combines feature extraction and classification into a single end-to-end optimisation problem (O'Shea et al., 2020).

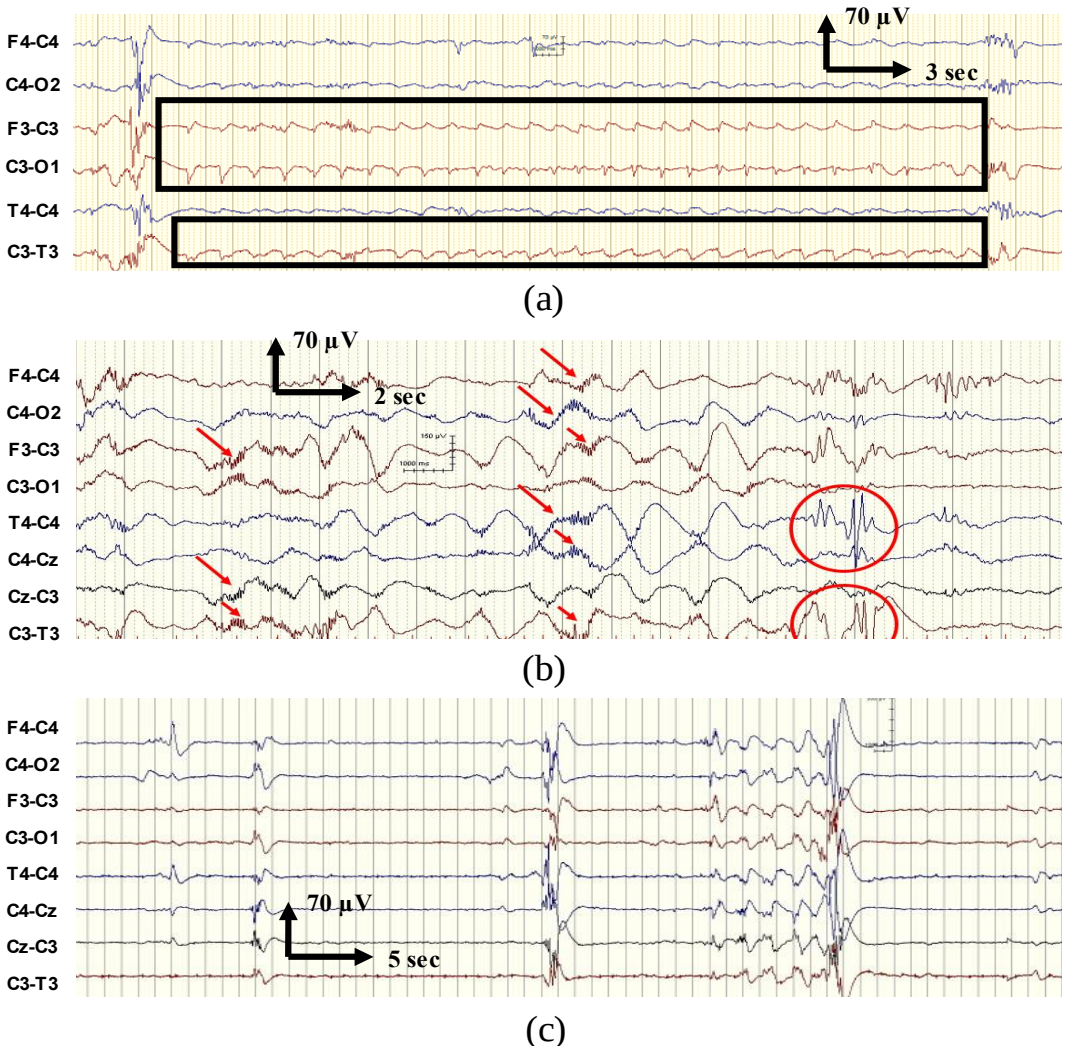


Figure 6.1 Examples of preterm background EEG patterns and a seizure event. (a) A representative example of preterm seizure showing a short 30-second lateralized seizure with rhythmic delta morphology and little evolution at 28 weeks GA, (b) runs of background semi-rhythmic delta activity with delta brush (arrows) and premature temporal theta transients (circles) at 31+5 weeks GA, (c) discontinuity at 24 weeks GA.

While there has been research in the area of automated classification of patterns in preterm background EEG (Stevenson et al., 2014; Koolen et al., 2015; Murphy et al., 2015; O'Toole et al., 2016, 2017), very few publications have investigated the development of automated seizure detection algorithms specifically for preterm infants. There are several important differences between term and preterm EEG which makes the development of preterm EEG seizure detection algorithms (P-SDA) challenging.

One challenging aspect to consider when developing a P-SDA is data availability. Recording multichannel EEG in the preterm is difficult due to the clinical condition of preterm infants and the physical limitations, such as the small head circumference. Specially trained staff are

required to record preterm EEG along with the correct equipment, and careful preparation (Lloyd et al., 2015). Good quality seizure and non-seizure data sets, which are required to train, test, and validate an algorithm for automated analysis, are therefore very difficult to obtain. The algorithmic solution to this problem must consider that comparatively small datasets will be available.

The characteristics of preterm EEG are very different to those of the term infant; this is another aspect to consider when designing a P-SDA. The features of preterm EEG that make P-SDA development more challenging include the intermittent background pattern (tracé discontinu) which consists of periods of low-voltage activity (inter-burst intervals) alternating with periods of short duration higher-voltage activity, known as bursts or spontaneous activity transients. Additionally, preterm seizures are usually focal/regional with less propagation than seizures of term neonates due to ongoing synaptogenesis and a relative paucity of inter-regional connectivity (Janáčková et al., 2016). The frequency of most preterm seizures lies in the delta band, but has been shown to evolve with age, and seizure morphology tends towards repetitive sharp waves, rather than spike and wave pattern (Janáčková et al., 2016; Patrizi et al., 2003). Preterm seizures are usually shorter in duration and demonstrate less evolution in terms of frequency and morphology during an ictal event. Figure 6.1 (a) shows an example of a preterm seizure with little spatial or frequency evolution. Given the complexity of preterm EEG, the set of seizure-descriptive features which were developed for a T-SDA machine learning model may not be optimal for preterm seizure detection.

Preterm EEG evolves rapidly as the infant matures and a thorough knowledge of the age appropriate maturational features of the EEG is required in order to identify seizures accurately (Pavlidis, et al., 2017b). Preterm EEG patterns change throughout gestation particularly in terms of voltage and length of discontinuous periods. Bursts of activity lengthen and inter-burst intervals progressively shorten during the preterm period and it has been suggested that this evolution from discontinuity to continuity represents a transition from primarily endogenously generated sporadic cortical activities in preterm to continuous sensory driven activity towards term (Vanhatalo & Kaila, 2006). In addition, specific features appear and disappear at different gestational ages and features such as asynchrony between hemispheres can be normal or abnormal depending on the gestational age (GA) (Pavlidis, 2017a, 2017b). There are several intermittent age appropriate maturational features such as delta brushes, and transient theta bursts over the temporal and occipital regions. The frequency content of the EEG also changes with GA; there is a predominance of slow delta activity at early preterm ages and an increase

in faster activities towards term (Jennekens et al., 2012). Figure 6.1 (b) and (c) provide examples of preterm background EEG patterns. The patterns in Figure 6.1 (b) and (c) show the characteristic changes that occur with GA. The question arises as to whether there is a need to employ models that are specific to GA, due to the evolution of preterm characteristics with GA; this work will look at both generalised and age specific classifier development. Given the scarcity of preterm data, it is important to consider efficient data usage when developing age specific preterm models, which are each naturally supported by a much smaller subset of the overall training data.

Previous work in the area of epileptic seizure detection in adult EEG has already shown that the algorithms developed for the adult population are not suitable for term neonatal EEG (Faul et al., 2005, 2009; Gotman et al., 1997; Liu et al., 1992). The clinical urgency and the outlined differences between term and preterm EEG need to be accommodated in the choice of data partition (train and test), system design, at the level of feature engineering and predictive modelling.

This study poses and addresses the following research questions:

i. Can the algorithms which were designed to detect seizures in term neonatal EEG also be used to detect seizures in preterm EEG?
ii. What are the steps to improve the model performance on preterm EEG?
iii. How can the existence of the term EEG seizure detection model and an abundance of term EEG data be leveraged to create an accurate preterm EEG seizure detection system with minimum preterm annotated EEG data in the minimum amount of time?

Two datasets containing annotated neonatal seizures were used in this study to develop and test the preterm seizure detection algorithms. Due to properties of the datasets, Preterm Dataset I, which has channel-specific seizure annotations, was used for training. Preterm Dataset II, which has no channel-specific seizure annotations but consists of long EEG recordings (575hours), was used for testing. More details on these datasets are provided in Chapter 2.

6.2 Testing term algorithm performance on Preterm Dataset II

Previous works have demonstrated the suitability of both machine learning based models and deep learning based models for the task of neonatal seizure detection (Ansari et al., 2018; O'Shea et al., 2020; Tapani et al., 2019; Temko et al., 2011). Despite the known and acknowledged differences between term EEG and preterm EEG the logical first step in developing a neonatal SDA for preterm EEG was to test the suitability of previously developed

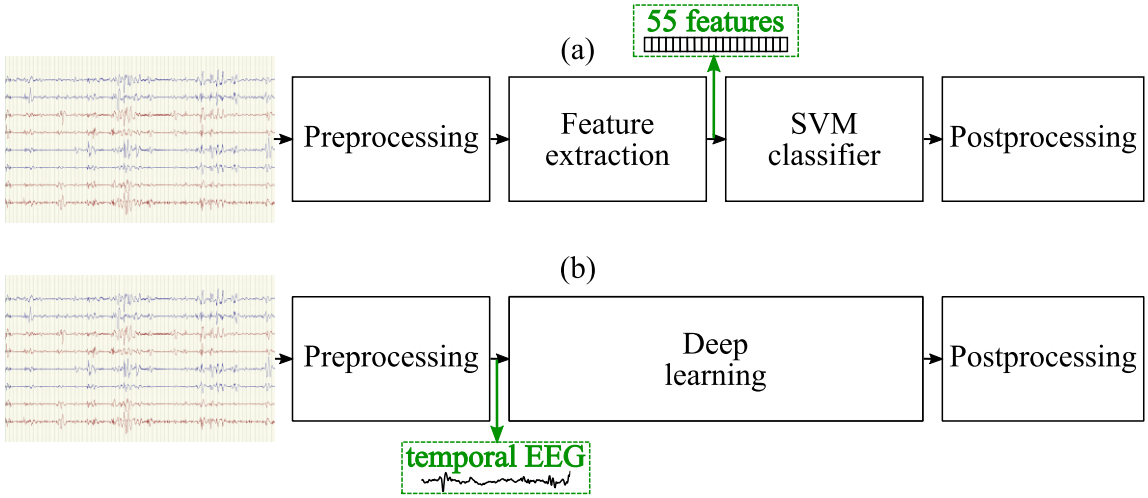


Figure 6.2 Comparison of the (a) SVM and (b) FCN algorithms. The SVM algorithm decomposes the EEG signal into features, the algorithm operates on one channel at a time, so channel specific annotations are required for training. The probabilistic outputs from the SVM algorithm are combined during the postprocessing using a maximum operation. The FCN is applied to multi-channel temporal EEG signals so it requires no feature extraction stage and there is one probabilistic output representing all channels.

T-SDAs. These experiments were intended to give an indication of the differences between the term seizure detection task and the preterm seizure detection task and to indicate a baseline performance level on Preterm Dataset II.

To establish a baseline performance, the previously developed Cork support vector machine (SVM) based T-SDA and the previously developed fully convolutional network (FCN) based T-SDA were tested on Preterm Dataset II (O'Shea et al., 2020; Temko et al., 2011). These algorithms represent two diverse approaches to detecting neonatal seizures; the Cork SVM algorithm relies on a set of extracted features and a machine learning classifier; the FCN algorithm predicts the seizure probability from temporal EEG windows using a convolutional neural network architecture. More details on these algorithms are given in Chapter 3 Section 3.3.2 and in Chapter 5. These algorithms were trained on a long dataset of multi-channel EEG from 18 term newborns with seizures secondary to hypoxic-ischemic encephalopathy (HIE) and one hour recordings from 54 term newborns who experienced HIE but did not have any EEG recorded seizures, more details on this Term Dataset I are given in Chapter 2 Section 2.6.1.

Figure 6.2 (a) shows the system level representation of the Cork SVM algorithm. This system is divided into four main stages: the preprocessing stage, the feature extraction stage, the SVM classification stage and the postprocessing stage. The preprocessing includes filtering, down sampling and segmentation into 8-second epochs. A set of 55 features are extracted from these

Table 6.1. The structure of each feature extraction block (a) and classification block (b) used in the deep learning T-SDA. The fully convolutional algorithm consists of three feature extraction blocks (a) followed by one classification block (b). This gives a total convolutional depth of 10 layers.

(a) Architecture of feature extraction building block

Layer Name	Feature Maps	Size	Stride	Activation
Convolution	32	3	1	ReLU
Convolution	32	3	1	ReLU
Convolution	32	3	1	ReLU
Batch Normalization				
Average Pooling		4	3	

(b) Architecture of classification building block

Layer Name	Feature Maps	Size	Stride	Activation
Convolution	2	3	1	ReLU
Global Pooling	-	-	-	
Output	-	-	-	Softmax

epochs; these features summarize the EEG activity using the time, frequency, and information theory domain characteristics. These features have been engineered specifically for term EEG and validated in a number of previous studies on term EEG, such as neonatal seizure detection (Greene et al., 2008; Temko et al., 2011; Thomas et al., 2010), grading term background EEG (Ahmed et al., 2016), neurological outcome prediction in term infants (Temko et al., 2015) and even adult seizure detection and prediction (Faul et al., 2009; Reuben et al., 2020). The extracted feature vectors are then fed to the SVM classification stage of the system where the inputs are assigned a probability of being a seizure. This probability is smoothed and postprocessed. Temko et al. have reported in detail on this algorithm (Temko et al., 2011, 2013).

Figure 6.2 (b) shows the system level representation of the deep learning algorithm. The deep learning system uses the same preprocessing and postprocessing stages as the SVM algorithm, but the feature extraction and classification stages are combined into one end-to-end deep learning network. The algorithm is applied to raw preprocessed EEG signals directly and the convolutional filters are used to extract data-driven information from 8-second windows of EEG. A series of hierarchical non-linear filtering operations extract a complex internal representation of the EEG, with the final convolutional layers performing the classification procedure, the architectural details are reported in Table 6.1. The designed fully convolutional architecture was optimised on term EEG (O'Shea et al., 2020).

The model was trained using a categorical cross-entropy loss function and a learning rate of 0.01. Three infants from the training set were removed and utilised as a patient independent validation set for early stopping. The AUC score on the validation set was calculated after each epoch - if the validation AUC score did not improve for eight training iterations the training

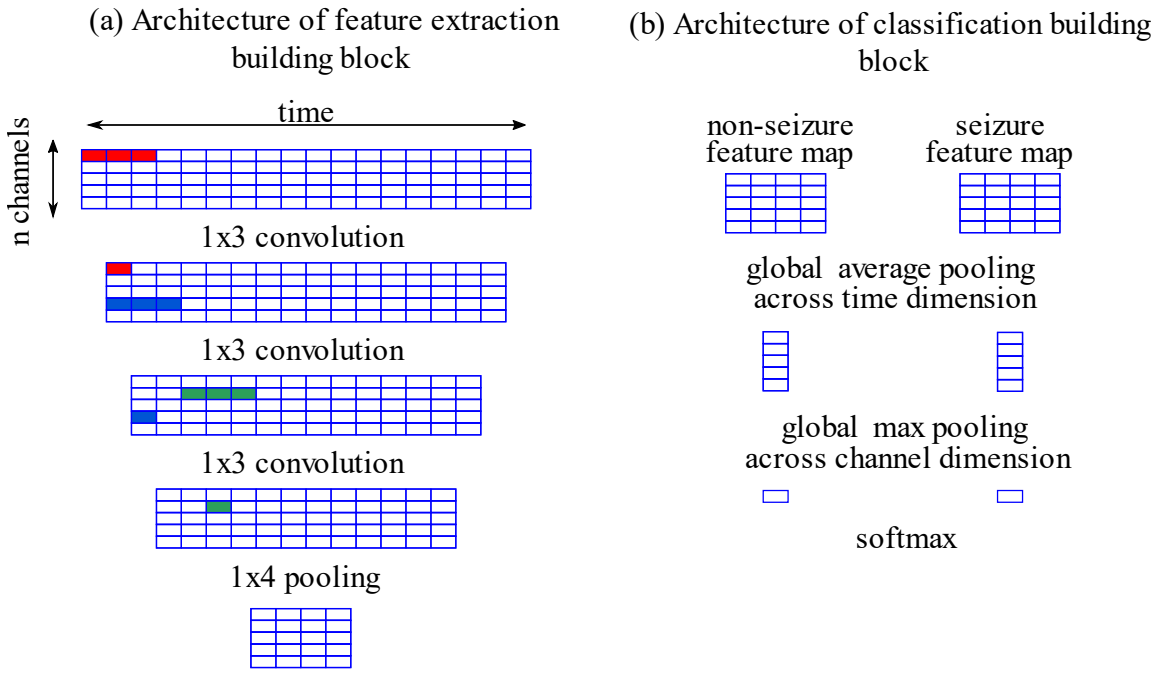


Figure 6.3 Illustration of the building blocks described in Table 6.1. Sample inputs are shown with time dimension across the x-axis and the number of EEG channels across the y-dimension.

was stopped and the network which resulted in the best validation AUC score was selected as the final network. This routine was repeated three times, randomly selecting a different set of infants for the validation set on each iteration. This resulted in three separate models, each based on a slightly different set of training infants. To fully utilise all three models, they are combined using an ensembling technique. The average of the probabilistic outputs from the three trained networks was taken to be the final probabilistic output at the inference stage.

An important aspect of this architecture is that it can be trained on multi-channel EEG signals - the seizure annotations do not therefore need to be channel specific. This property meant that the FCN T-SDA algorithm was trained on a dataset of over 800 hours of EEG (O'Shea et al., 2020). In contrast to this, the SVM T-SDA was trained on a subset of around 2% of the same

Table 6.2 Performance of the T-SDAs on Preterm Dataset II in AUC (%). The first column reports the performance of both algorithms on the Term Data I; this is the dataset which was utilised to develop and train the T-SDAs. The second column shows the T-SDA performances on the entire Preterm Dataset II. The remaining columns show the performance measured on subsets of the dataset relating to each GA group.

T-SDA	Term Dataset I	Preterm Dataset II	Group 1	Group 2	Group 3
SVM	96.6	88.3	93.6	87.1	86.8
FCN	98.5	93.3	91.8	96.0	93.8

term EEG dataset because it required channel-specific seizure annotations during the training stage. The generation of channel-specific seizure annotations is a laborious process and typically only a subset of seizure events in a dataset are annotated in this way.

Table 6.2 shows the results of the two T-SDAs when tested on Preterm Dataset II. It can be seen that the SVM T-SDA drops its performance from an AUC of 96.6% on Term Dataset I to an AUC of 88.3% on Preterm Dataset II, which is equal to an 8.3% absolute drop in performance. For the FCN-based T-SDA, a smaller decrease in AUC of 5.2% absolute is observed, dropping from an AUC of 98.5% to an AUC of 93.3%. Table 6.2 also presents algorithm performance for different GA groups within the dataset; the 16 infants in Preterm Dataset II were divided into three subgroups according to the gestational ages of the infants: Group 1 (GA < 26 weeks), Group 2 (26 weeks ≤ GA ≤ 29 weeks) and Group 3 (GA > 29 weeks). It is known that the GA of a preterm has a large effect on EEG morphology, analysing performance across GA groups gives a clearer insight into model suitability.

Previous works have proposed several algorithms to detect seizures in full term EEG. It has been shown that both the approaches that rely on hand-crafted features and the approaches that are based on deep learning achieve reliable and robust results when trained and tested on term EEG. This work however illustrates that the performance of these approaches falls short when tested on preterm EEG. The results in Table 6.2 indicate that the SVM T-SDA, which is based on a set of hand-crafted features, in particular, is not suitable for use as a preterm seizure

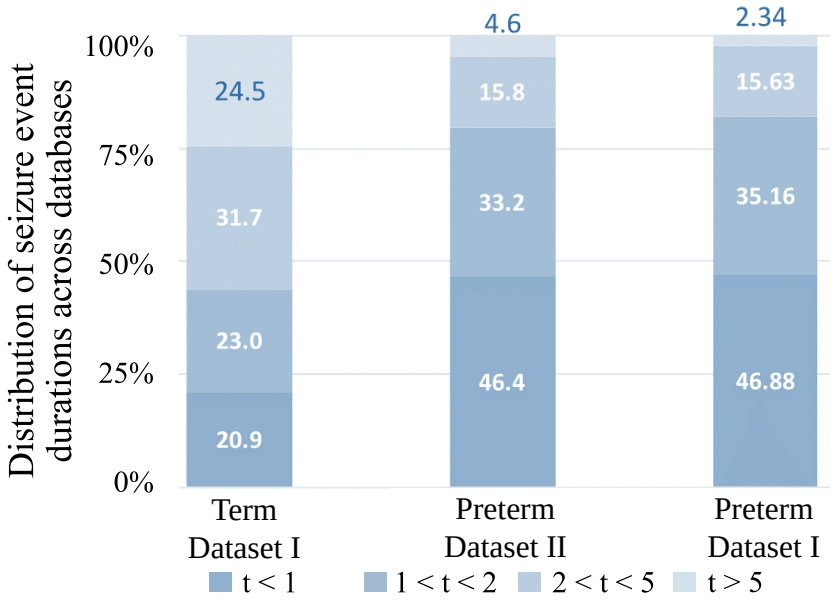


Figure 6.4 Distribution of seizure events according to their duration in the term and preterm datasets. t represents the duration of the seizure event in minutes.

detection tool. Various ways to improve the performance of SDAs designed for term EEG are explored in this work and assessed on Preterm Dataset II.

Figure 6.4 compares the duration of seizure events in the preterm datasets utilised in this work against Term Dataset I, which was used to develop two term EEG SDAs (O'Shea et al., 2020; Temko et al., 2011). It can be seen that preterms have double the proportion of short seizures, approximately 46% (less than 1-minute duration), when compared to the term infants. Moreover, preterm datasets have the smallest proportion of seizures with duration of greater than 5 minutes. While the distributions of seizure durations in the term and preterm populations differ, the two preterm datasets, Preterm Dataset I and Preterm Dataset II, have similar seizure event statistics. The SVM T-SDA algorithm utilised in this work has been proven to be most sensitive to longer seizure events, the algorithm performance was reduced when tested on short duration seizures (Mathieson et al., 2016). This characteristic may explain the decrease in performance of then SVM T-SDA when tested on Preterm Dataset II.

6.3 Algorithm retraining: Preterm Dataset I with the term learning framework

The decrease in T-SDA performance for both the SVM and FCN classifiers from the term domain to the preterm domain, illustrates that morphology of these physiological signals is different. The results in Table 6.2 indicate that the signal characteristics used by both algorithms to distinguish seizures in the term domain are not optimal for the preterm domain. This finding motivated the development of preterm specific SDAs. Preterm Dataset I was employed as a training dataset for this task and Preterm Dataset II was again employed as the testing dataset.

The simplest way to build a P-SDA is to re-use an existing T-SDA algorithm architecture but to retrain it on preterm EEG data. In this manner, the same preprocessing, feature extraction, modelling and postprocessing stages can be employed. In experiments both the SVM and FCN SDA architectures were trained using Preterm Dataset I to develop P-SDAs. Due to the need for channel-specific seizure annotations the SVM based algorithm was trained on the subset of seizure events which had channel-specific seizure annotations. Preterm Dataset I consists of 133 seizure events, 73 of these events have at least a subset of the total event annotated with strong labels. Conversely, the FCN algorithm was trained on all the available multi-channel preterm EEG. Preterm Dataset I consists of recordings with 14 bipolar EEG channels and Preterm Dataset II consists of recordings with 8 bipolar EEG channels, the FCN algorithm is invariant to the number of input channels due to the fully convolutional design so it can be

Table 6.3 Performance of the P-SDAs, which were trained using Preterm Dataset I, when tested on Preterm Dataset II in AUC (%). The first column reports the P-SDA performances on the entire Preterm Dataset II. The remaining columns show the performance measured on subsets of the dataset relating to each GA group.

P-SDA	Preterm Dataset II	Group 1	Group 2	Group 3
SVM	89.7	87.1	71.8	96.2
FCN	93.5	94.4	84.3	92.1

adapted to train and test on these variable sized input data representations. This experiment will show how suitable the system choices (engineered features, FCN architecture, etc) which were made for term EEG scenario are for classification of preterm EEG.

The performances of the SVM and FCN algorithms when trained using Preterm Dataset I are shown in Table 6.3. When retrained on a dataset of preterm EEG, the performance of both the SVM and FCN SDAs on the entire Preterm Dataset II marginally improves with respect to the T-SDAs, with a larger improvement observed for the SVM P-SDA. However, the GA group performances for both algorithms do not show this linear improvement.

6.3.1 GA-specific algorithm development

It can be seen from Table 6.2 and Table 6.3 that the performance of both the T-SDAs and P-SDAs differ substantially across the GA groups. While overall results of retraining using preterm data show marginal improvement over their corresponding T-SDAs counterparts, the distribution of scores in the three groups between P-SDAs and T-SDAs vary widely. These results indicate that the same algorithm may not be optimal for detecting seizures in infants from different GA groups due to the variations in EEG morphology and seizure manifestation. These insights motivated the development of GA specific classifiers which are designed to optimise the seizure detection rate for a GA specific subset of the preterm datasets.

Preterm Dataset I contains EEG data from 17 infants; four of these infants fall into GA group 1, five of these infants fall into GA group 2, and the remaining eight infants fall into GA group 3. In order to train GA specific SVM P-SDA and FCN P-SDA, these subsets of the training dataset are utilised as the training data. The SVM training routine was not altered for these GA specific experiments. For the FCN training routine, a validation dataset was again utilised to perform early stopping. When the validation AUC stops increasing for 8 training iterations the training routine ceases and the model which produced the highest validation AUC is taken to be the trained model. Due to data limitations, in the GA-specific model training routine the

Table 6.4 Performance of the GA specific P-SDAs, which were trained using GA specific subsets of Preterm Dataset I, when tested on the matched GA specific subsets of Preterm Dataset II in AUC (%).

GA specific P-SDA	Group 1	Group 2	Group 3
SVM	82.6	91.6	90.7
FCN	85.8	76.0	74.3

validation set consists of just two infants and both of these infants come from the target GA group.

Table 6.4 presents the GA group specific results of the GA specific P-SDA algorithm development experiments. To evaluate the effect of the GA conditional changes in preterm EEG, the performance of the SDAs was evaluated for each GA group separately and compared. It can be seen from Table 6.3 and Table 6.4 that the performance of the SVM P-SDAs on individual GA groups varies widely depending on the GA. No clear pattern of improvement was observed when three different GA-specific SVM classifiers were developed by segmenting Preterm Dataset I into groups based on GA. This decreases the overall AUC across concatenated data and indicates a lack of robustness in the algorithm.

Comparing the FCN P-SDA results in Table 6.3 and Table 6.4 indicates that the FCN algorithm is detrimentally affected by the segmentation of Preterm Dataset I into GA subsets. This degradation in FCN performance could be attributed to the small amount of data available in each GA subgroup of Preterm Dataset I. While the small improvement observed for the SVM P-SDA (when the SVM T-SDA was retrained with Preterm Dataset I) was mainly limited by the reutilisation of features which were designed for term infants, the diminished performance that was observed in the FCN P-SDA was probably influenced by the size of the preterm training data, which was not sufficiently large to learn the complexity and variation in preterm EEG.

6.4 Modified FCN algorithm training: GA-specific retraining and transfer learning

The FCN algorithm training framework utilised in Section 6.3 for simple P-SDA development and GA-specific P-SDA development was established in Chapter 5 where over 800 hours of multi-channel training data were available. The preterm training data available to train P-SDAs are less than 24 hours in duration. Given the scarcity of the training data, segmenting data into GA subgroups comes at a cost of further reducing the training data for each model. SVMs are known to perform well with limited amounts of training data and thus for the SVM-based SDAs the training process for the GA-specific experiments is unchanged. However, the FCN is

known to require large amounts of training data to be accurate, and the scarcity of available training data in each GA-specific experiment poses a real problem for the FCN-based SDAs. To overcome the limited training data availability when training the GA-specific FCN SDAs the training process has been modified in three ways:

i) The FCN GA-specific P-SDA was trained using transfer learning (TL) instead of retraining from a random initialisation. In particular, the network was initialised with weights from one of three T-SDA models which were trained on large amounts of term EEG. Three models were trained during the development of the T-SDA algorithm, each was trained using a random subset of the training infants and the remaining infants were utilised for model validation, for each of the three models trained in the P-SDA algorithms a different one of the three T-SDA models was employed for weight initialisation. The data from each target GA group were used to fine-tune the convolutional layers.

ii) Layer-wise adaptive learning rate scaling (LARS) was utilised (You et al., 2017). LARS extends the stochastic gradient descent with momentum to determine a learning rate per layer. This is done to account for the potential variation of parameter (weights and biases) and gradient magnitudes in each layer during the transfer learning stage. In typical standard gradient descent updates, the weight update is $w_{ij}^{l} \leftarrow w_{ij}^{l} - \eta \nabla w_{ij}^{l}$, where w_{ij}^{l} is a weight in the l^{th} layer, ∇w_{ij}^{l} is the gradient of that weight with respect to the loss function and η is the learning rate. LARS scales the weight update by the layer-wise ratio of the l_2-norm of the parameters and the l_2-norm of the gradients.

$$w_{ij}^{l} \leftarrow w_{ij}^{l} - \eta \frac{\|w^{l}\|_2}{\|\nabla w^{l}\|_2} \nabla w_{ij}^{l} \tag{6.1}$$

iii) Instead of excluding data from GA group specific algorithms during training all the available training data are included in the FCN TL experiment. Each training sample is weighted based on one of the three membership functions shown in Figure 6.5. In particular, the training samples that belong to the chosen GA group are given a weight of 1.0 in the loss; the remaining data are given a weight which linearly decays depending on the difference between their GA and the GA threshold of the group for which the P-SDA is trained.

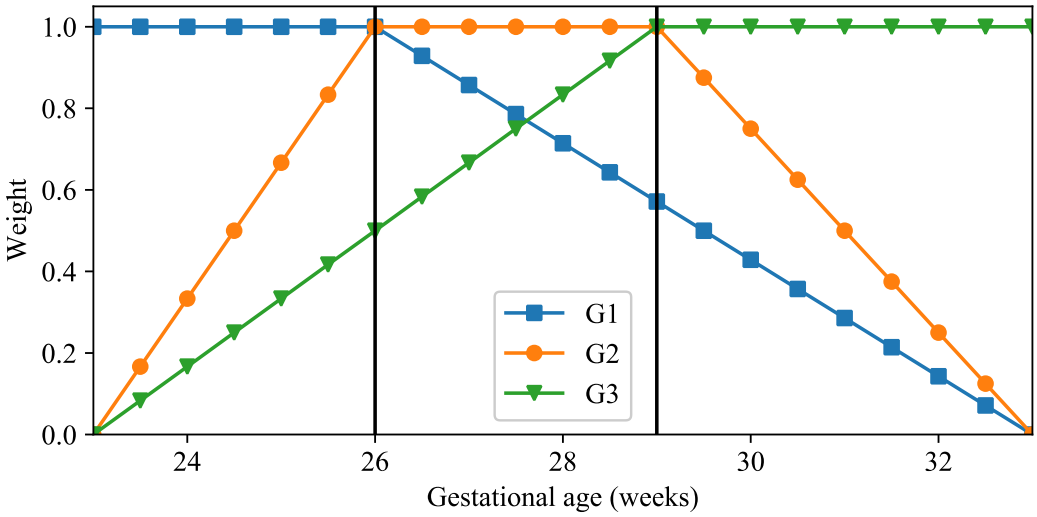


Figure 6.5 The weighting applied to each preterm infant's GA during training of the GA specific FCN classifier. These weights are utilised during the training of the GA specific algorithm to ensure that all of the training data were used despite the developed classifiers being age specific.

The results in Table 6.5 indicate that modifications i, ii, and iii contribute to a more stable improvement in classifier performance. This modified training framework is more appropriate for the preterm SDA development domain; where data is scarce, there is a wide variation in EEG morphology, and a previously trained robust classifier is available for TL. When comparing the GA-specific retraining with the simple retraining results in Section 6.3, the SVM P-SDA did not improve in all three groups, whereas the GA-specific TL FCN P-SDA improved across all three groups; improving from an AUC 93.5% to 95.0% when tested across all infants concatenated in Preterm Dataset II.

Deep learning based algorithms require large amounts of training data. For instance, the term algorithm was trained on a dataset of over 800 hours of multi-channel EEG. Preterm Dataset I has less than 24 hours of EEG, a relatively small dataset for deep learning algorithm development. Dividing this dataset into three GA specific groups resulted in even smaller datasets to train each GA-specific FCN P-SDA. To overcome the constraints imposed by data scarcity, three different classifiers were trained using the complete Preterm Dataset I as

Table 6.5 Performance of the GA specific TL FCN P-SDA, which was trained on data from all infants in Preterm Dataset I, where each infant was weighted based on their GA, in AUC (%). This algorithm benefits from the utilisation of TL from the term trained FCN T-SDA.

GA specific TL P-SDA	Preterm Dataset II	Group 1	Group 2	Group 3
FCN	95.0	95.8	93.8	93.4

explained in Section 6.4. Additionally, the GA specific DL P-SDA training also benefitted from the utilization of transfer learning; a deep learning technique which allows for the initialization of algorithms using pre-trained network weights borrowed from the FCN T-SDA.

It is worth noting that the FCN P-SDA was trained using seizure annotations that contained no information about the seizure locations. Preterm seizures are known to start and remain more focal than term seizures with less propagation (Janáčková et al., 2016; Patrizi et al., 2003). The FCN P-SDA takes multi-channel EEG as input and a single seizure or non-seizure label as target. The training dataset in this work consists of 14 channels of EEG, therefore the seizure to non-seizure ratio could be as poor as 1-to-13 for an FCN P-SDA training epoch. But the ability to learn from these weakly labelled examples means that all the data in Preterm Dataset I were available to train the algorithm.

6.5 Performance improvement: a mixture of experts

The FCN models outperform the SVM SDAs on both term and preterm datasets, reaching an AUC of 98.5% for term and 95% for preterm as compared to 96.6% and 89.7% for SVM term and preterm SDAs, respectively. Interestingly, when retrained on Preterm Dataset I from scratch and with TL, both SVM and FCN P-SDAs do not improve the performance for all groups. In fact, the best performance for Group 3 (GA > 29 weeks) is obtained with the SVM preterm model (SVM P-SDA), for Group 2 (26 weeks ≤ GA ≤ 29 weeks) with the deep learning term model (FCN T-SDA), for Group 1 (26 weeks ≤ GA ≤ 29 weeks) and overall for all three groups with GA-specific transfer learning (FCN P-SDA TL GA-specific). Taking the best model for each group and using the resultant meta-classifier will inevitably improve the performance on the observed dataset. However, this will come at a cost of using the same data to select the models and assess the performance of the models which will invalidate the latter. The choice of models must be made without observing the Preterm Dataset II performance.

The final stage of development in this work was a fusion of various models and selection of the best combination on the validation dataset which was formed from Preterm Dataset I. Algorithm fusion was employed to exploit the diverse range of models that were generated to solve the preterm seizure detection challenge and to increase the power of the developed P-SDA. To assess this hypothesis, two simple fusion schemes were tested using either the weighted arithmetic mean (6.2) or the weighted geometric mean (6.3) over the probabilistic output of the classifier, *P*, which are defined as follows, for $\alpha_i \geq 0, \sum_i \alpha_i = 1$:

$$P_{fusion\_arithmetic\_mean} = \sum_i \alpha_i P_i \tag{6.2}$$

$$P_{fusion\_geometric\_mean} = \prod_i P_i^{\alpha_i} \tag{6.3}$$

The two classification algorithms selected for fusion in the experiment were the GA-specific TL FCN P-SDA and the FCN T-SDA. Overall, it can be seen the FCN SDAs outperformed the SVM SDA both on term and preterm EEG. These specific FCN SDAs were selected for fusion based on their performance in AUC across all three GA groups and their perceived diversity. For a combination of two classifiers, Equation 6.2 can be rewritten as $P_{fusion\_a\_mean} = \alpha P_{T-SDA} + (1-\alpha)P_{P-SDA}$. and Equation 6.3 can be rewritten as $P_{fusion\_g\_mean} = P_{T-SDA}^{\alpha} P_{P-SDA}^{1-\alpha}$. To select an appropriate α value the validation set, which was used for early-stopping during FCN training, was reutilised.

Through testing a range of α values on the validation infants an optimal blending ratio was found based on the calculated AUC, and this was then tested on Preterm Dataset II. Figure 6.6 shows the AUC obtained on the validation set (a subset of the training data) as a function of the chosen weight for both geometric and arithmetic mean ensembling methods. It can be seen

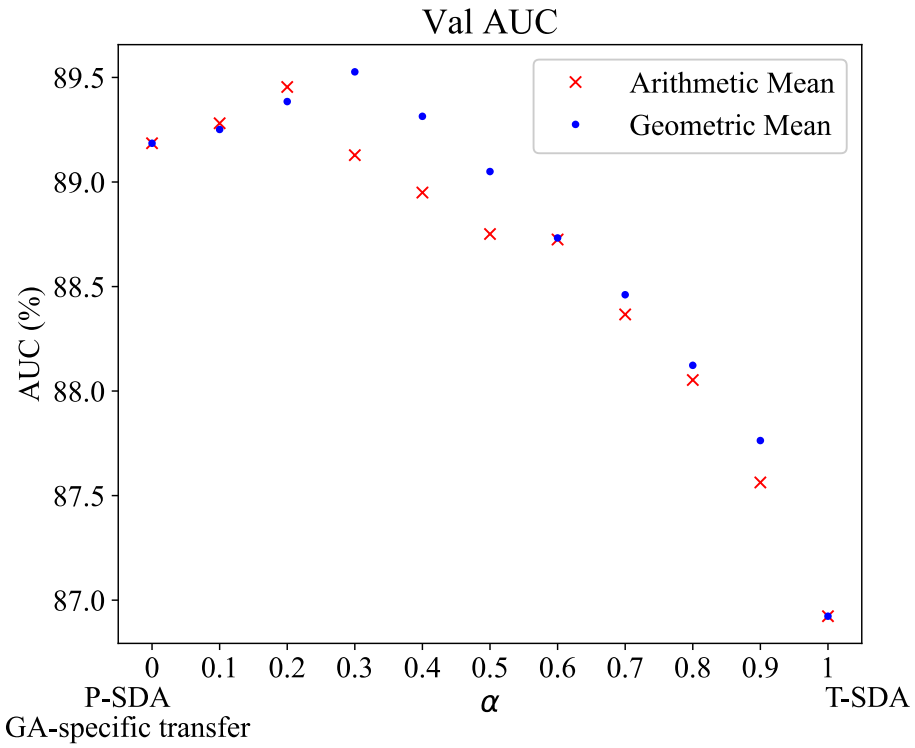


Figure 6.6 The AUC performance on the validation dataset as a function of α. The performance reported here is for the two held-out preterm infants, which were used for early stopping. The leftmost point represents the FCN P-SDA with GA specific transfer learning (α = 0), the rightmost point represents the FCN T-SDA (α = 1). Validation AUC is taken without postprocessing and thus can be less than test performance.

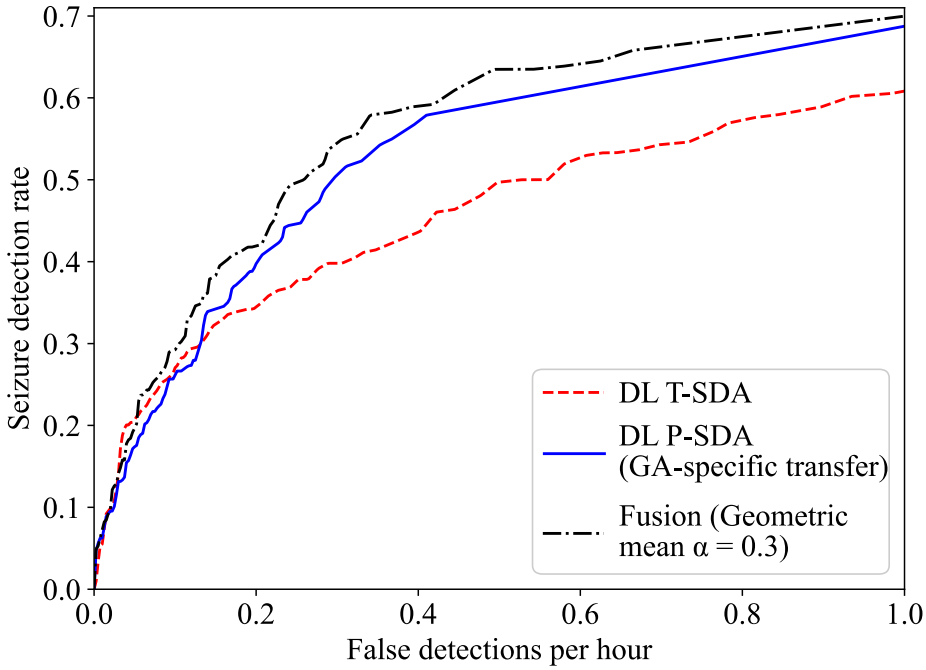


Figure 6.7 The curve of seizure detection rate vs the number of false detections per hour; each point on the curve represents a different threshold decision value. The fusion with geometric mean reaches an overall AUC of 95.4% across all infants in Preterm Dataset II.

from Figure 6.6 that the geometric mean of probabilities given by FCN T-SDA and FCN P-SDA TL GA-specific gives the highest performance on the validation data when combined with a weight of 0.7 given to P-SDA and a weight of 0.3 given to T-SDA.

While the AUC metric represents an evaluation of the epoch-based accuracy, which is an important tool for engineering development, the full picture of an algorithm's clinical utility can be seen through event-based metrics such as the seizure detection rate and the number of false detections per hour. Figure 6.7 shows the results of clinical metrics, seizure detection rate and FD/h, for the best combination of the SDA systems, and the results of the individual SDA systems used in the ensemble (FCN T-SDA and FCN P-SDA TL GA-specific). The method of fusion and the best weighting of each classifier were found on the validation data (previously utilised for early stopping in FCN systems). Considering a threshold value which gives one false detection every four hours the term algorithm (FCN T-SDA) detects 37.6% of seizures, the GA specific preterm algorithm with transfer learning (FCN P-SDA TL GA-specific) detects 44.6%, and the fusion of these two classifiers by geometric mean, with 0.7 and 0.3 weights for preterm and term algorithms, respectively, results in 49.7% of seizure events being detected. Importantly, this mixture of experts outperforms the best DL P-SDA GA-specific transfer for the entire range of operating points. The final best overall AUC for this fusion of classifiers

was 95.4% on Preterm Dataset II. It is important to stress that the reported performance is not the best achievable on this dataset, but it is instead the performance that is anticipated through model selection on the validation data.

6.6 Conclusion

The work in this chapter represents the first time an automated seizure detection system has been developed specifically for preterm EEG. This work has contributed three important messages. First, the study showed that algorithms developed for term EEG exhibited performance degradation when tested on preterm EEG and therefore were not suitable to be used for seizure detection in preterm EEG. Second, no substantial increase in algorithm performance was observed when the systems were simply retrained on preterm EEG, this is potentially due to the scarcity of available training data and can attributed to the variation of EEG morphology with GA. Finally, it was demonstrated that the lengthy process of manual feature engineering could be avoided, and an accurate SDA can be developed, with minimal preterm data availability.

In this work the Cork SVM classifier was trained to detect seizures from EEG windows which were decomposed into a large set of engineered features. Results showed that this machine learning algorithm was not suitable for preterm seizure detection. There was no substantial increase in algorithm performance when it was retrained on preterm EEG. This potentially indicates that the features utilised in the machine learning algorithm are not suitable for the detection of preterm seizures and that the algorithm cannot deal with the larger level of noise in the ground truth due to the high inter-observer variability in preterm datasets.

This work also proposed utilizing an end-to-end deep learning algorithm which was previously validated on term EEG. The deep learning approach does not require an engineered feature set and it can train from weakly labelled seizure events, unlike the SVM classifier. The proposed novel end-to-end deep learning approach utilises three GA-specific models and overcomes limited data availability by 1) weighting data from all GA subgroups 2) utilising TL from the term model 3) not requiring per-channel annotations and therefore learning from the more readily available weak labels.

The term FCN algorithm and the GA specific preterm FCN algorithm represented the best performing models on the preterm test dataset. The fusion of these algorithms using a routine selected on a validation dataset further improved the performance of the designed preterm seizure detection system. The performance obtained by combining the best performing

algorithms reached an AUC score of 95.4%; while detecting nearly half of preterm neonatal seizure events at a clinically acceptable cost of one false seizure detection every 4 hours

These results indicate that a robust reliable preterm seizure detection algorithm has been trained without any additional feature engineering by utilizing deep learning. The results obtained using deep learning allow for its practical application in NICUs for the automatic detection of preterm seizures. To improve the deep learning classifier further a larger dataset of preterm EEGs with annotated seizures could be acquired, but this data would not require channel-specific seizure annotations, thus reducing the workload on clinical staff.

6.7 References

Ahmed, R., Temko, A., Marnane, W., Lightbody, G., & Boylan, G. (2016). Grading hypoxic–ischemic encephalopathy severity in neonatal EEG using GMM supervectors and the support vector machine. *Clinical Neurophysiology*, *127*(1), 297–309.

Ahmed, R., Temko, A., Marnane, W. P., Boylan, G., & Lightbody, G. (2017). Exploring temporal information in neonatal seizures using a dynamic time warping based SVM kernel. *Computers in Biology and Medicine*, *82*, 100–110.

Ansari, A. H., Cherian, P. J., Dereymaeker, A., Matic, V., Jansen, K., De Wispelaere, L., Dielman, C., Vervisch, J., Swarte, R. M., Govaert, P., Naulaers, G., De Vos, M., & Van Huffel, S. (2016). Improved multi-stage neonatal seizure detection using a heuristic classifier and a data-driven post-processor. *Clinical Neurophysiology*, *127*(9), 3014–3024.

Ansari, A. H., Cherian, P. J., Caicedo, A., Naulaers, G., De Vos, M., & Van Huffel, S. (2018). Neonatal Seizure Detection Using Deep Convolutional Neural Networks. *International Journal of Neural Systems*, *29*(04), 1850011.

Bogaarts, J. G., Gommer, E. D., Hilkman, D. M. W., van Kranen-Mastenbroek, V. H. J. M., & Reulen, J. P. H. (2016). Optimal training dataset composition for SVM-based, age-independent, automated epileptic seizure detection. *Medical & Biological Engineering & Computing*, *54*(8), 1285–1293.

Buraniqi, E., Sansevere, A. J., Kapur, K., Bergin, A. M., Pearl, P. L., & Loddenkemper, T. (2017). Electrographic Seizures in Preterm Neonates in the Neonatal Intensive Care Unit. *Journal of Child Neurology*, 0883073817713918.

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

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

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

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

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

Janáčková, S., Boyd, S., Yozawitz, E., Tsuchida, T., Lamblin, M.-D., Gueden, S., & Pressler, R. (2016). Electroencephalographic characteristics of epileptic seizures in preterm neonates. *Clinical Neurophysiology*, *127*(8), 2721–2727.

Jennekens, W., Niemarkt, H. J., Engels, M., Pasman, J. W., van Pul, C., & Andriessen, P. (2012). Topography of maturational changes in EEG burst spectral power of the preterm infant with a normal follow-up at 2 years of age. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *123*(11), 2130–2138.

Khazipov, R., Sirota, A., Leinekugel, X., Holmes, G. L., Ben-Ari, Y., & Buzsáki, G. (2004). Early motor activity drives spindle bursts in the developing somatosensory cortex. *Nature*, *432*(7018), 758–761.

Koolen, N., Dereymaeker, A., Räsänen, O., Jansen, K., Vervisch, J., Matic, V., De Vos, M., Naulaers, G., Van Huffel, S., & Vanhatalo, S. (2015). Data-driven metric representing the maturation of preterm EEG. *Proceedings of Annual International Conference of the IEEE on Engineering in Medicine and Biology Society (EMBC)*, 1492–1495.

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

Lloyd, R., Goulding, R., Filan, P., & Boylan, G. (2015). Overcoming the practical challenges of electroencephalography for very preterm infants in the neonatal intensive care unit. *Acta Paediatrica*, *104*(2), 152–157.

Lloyd, R., O'Toole, J. M., Pavlidis, E., Filan, P. M., & Boylan, G. B. (2017). Electrographic Seizures during the Early Postnatal Period in Preterm Infants. *The Journal of Pediatrics*, *187*, 18-25.e2.

Mathieson, S. R., Stevenson, N. J., Low, E., Marnane, W. P., Rennie, J. M., Temko, A., Lightbody, G., & Boylan, G. B. (2016). Validation of an automated seizure detection algorithm for term neonates. *Clinical Neurophysiology*, *127*(1), 156–168.

Milh, M., Kaminska, A., Huon, C., Lapillonne, A., Ben-Ari, Y., & Khazipov, R. (2007). Rapid cortical oscillations and early motor activity in premature human neonate. *Cerebral Cortex*, *17*(7), 1582–1594.

Murphy, K., Stevenson, N. J., Goulding, R. M., Lloyd, R. O., Korotchikova, I., Livingstone, V., & Boylan, G. B. (2015). Automated analysis of multi-channel EEG in preterm infants. *Clinical Neurophysiology*, *126*(9), 1692–1702.

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

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

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

Okumura, A., Hayakawa, F., Kato, T., Itomi, K., Maruyama, K., Kubota, T., Suzuki, M., Kidokoro, H., & Watanabe, K. (2008). Ictal electroencephalographic findings of neonatal seizures in preterm infants. *Brain and Development*, *30*(4), 261–268.

O'Shea, A., Lightbody, G., Boylan, G., & Temko, A. (2020). Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture. *Neural Networks*, *123*, 12–25.

O'Shea, A., Lightbody, G., Boylan, G., & Temko, A. (2017). Neonatal seizure detection using convolutional neural networks. *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6.

O'Toole, J., Boylan, G., Vanhatalo, S., & Stevenson, N. (2016). Estimating functional brain maturity in very and extremely preterm neonates using automated analysis of the electroencephalogram. *Clinical Neurophysiology*, *127*(8), 2910–2918.

O'Toole, J., Boylan, G. B., Lloyd, R. O., Goulding, R. M., Vanhatalo, S., & Stevenson, N. J. (2017). Detecting bursts in the EEG of very and extremely premature infants using a multi-feature approach. *Medical Engineering & Physics*, *45*, 42-50.

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

Pavlidis, E., Lloyd, R. O., & Boylan, G. B. (2017a). EEG - A Valuable Biomarker of Brain Injury in Preterm Infants. *Developmental Neuroscience*, *39*(1–4), 23–35.

Pavlidis, E., Lloyd, R. O., Mathieson, S., & Boylan, G. B. (2017b). A review of important electroencephalogram features for the assessment of brain maturation in premature infants. *Acta Paediatrica (Oslo, Norway: 1992)*, *106*(9), 1394–1408.

Reuben, C., Karoly, P., Freestone, D. R., Temko, A., Barachant, A., Li, F., Titericz, G., Lang, B. W., Lavery, D., Roman, K., Broadhead, D., Jones, G., Tang, Q., Ivanenko, I., Panichev, O., Proix, T., Náhlík, M., Grunberg, D. B., Grayden, D. B., Kuhlmann, L. (2020). Ensembling crowdsourced seizure prediction algorithms using long-term human intracranial EEG. *Epilepsia*, *61*(2), e7–e12.

Shellhaas, R. A. (2019). Chapter 17—Seizure classification, etiology, and management. In L. S. de Vries & H. C. Glass (Eds.), *Handbook of Clinical Neurology* (Vol. 162, pp. 347–361). Elsevier.

Stevenson, N. J., Palmu, K., Wikström, S., Hellström-Westas, L., & Vanhatalo, S. (2014). Measuring brain activity cycling (BAC) in long term EEG monitoring of preterm babies. *Physiological Measurement*, *35*(7), 1493.

Tapani, K. T., Vanhatalo, S., & Stevenson, N. J. (2019). Time-Varying EEG Correlations Improve Automated Neonatal Seizure Detection. *International Journal of Neural Systems*, *29*(04), 1850030.

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

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

Temko, A., & Lightbody, G. (2016). Detecting Neonatal Seizures With Computer Algorithms. *Journal of Clinical Neurophysiology*, *33*(5), 394–402.

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

Thomas, E. M., Temko, A., Lightbody, G., Marnane, W. P., & Boylan, G. B. (2010). Gaussian mixture models for classification of neonatal seizures using EEG. *Physiological Measurement*, *31*(7), 1047–1064.

Thomas, E. M., Temko, A., Lightbody, G., Marnane, W. P., & Boylan, G. B. (2011). Advances in Automated Neonatal Seizure Detection. In A. E. Ruano & A. R. Várkonyi-Kóczy (Eds.), *New Advances in Intelligent Signal Processing* (pp. 93–113). Springer.

Vanhatalo, S., & Kaila, K. (2006). Development of neonatal EEG activity: From phenomenology to physiology. *Seminars in Fetal and Neonatal Medicine*, *11*(6), 471–478.

You, Y., Gitman, I., & Ginsburg, B. (2017). Large Batch Training of Convolutional Networks. *ArXiv:1708.03888 [Cs]*.

Chapter 7: Technology translation for effective clinical decision support

Deep learning has been at the forefront of research and development in physiological signal processing in recent years (Faust et al., 2018). Translating this research to the clinical domain is challenging; there are many regulatory and clinical constraints which must be navigated before an algorithm can be adopted into clinical usage. This chapter will explore two aspects of deep learning algorithm translation which may help to bridge the gap between academic research and clinical realisation. The first aspect which will be addressed in this chapter is algorithm interpretability and the importance of understanding decision making factors in medical settings. The second aspect of algorithm translation that will be explored in this chapter are the considerations and adaptations which may be necessary to implement deep learning algorithms as part of portable hand-held EEG recording and interpretation devices.

Algorithm transparency and interpretable decision making are important when it comes to automatic decision support systems in medicine (Shortliffe and Sepúlveda, 2018). Deep learning is considered to be a black-box, but many researchers have focused on developing methods to understand deep learning algorithms (Choo and Liu, 2018; Simonyan and Zisserman, 2015). One way of probing deep learning algorithms focuses on saliency mapping for individual input examples. This class of methods seeks to understand the features of a specific input example which contribute to the label assigned by a deep learning algorithm. The second method of algorithm understanding focuses on generating examples which activate certain filters in the hidden layers of a trained deep learning architecture. This allows researchers to visualise the hierarchical data-driven features which the network utilises to make decisions. Recent work in EEG sleep stage scoring has demonstrated the importance and feasibility of interpreting trained deep learning algorithms (Andreotti et al., 2018; Vilamala et al., 2017). These works show that algorithm interpretability can become part of the research and development process for automated EEG analysis systems. Many previous works have focused on the visualisation and understanding of deep learning algorithms with 2-dimensional inputs. EEG is a temporal signal and neurophysiological experts interpret signal characteristics in the time domain; the ability to interpret deep learning algorithms with 1-dimensional temporal inputs may provide more value to clinical experts who seek to understand the decisions made by deep learning algorithms.

Neurophysiologists have highlighted the need for automated neonatal seizure detection algorithms to provide real-time clinical decision support (Ramantani et al., 2019; Vanhatalo, 2011). Two algorithms have been evaluated as potential automated seizure detection algorithms for neonatal populations (Navakatikyan et al., 2006; Temko et al., 2011). Navakatikyan's algorithm is only approved for post-hoc EEG review (Temko and Lightbody, 2016). The ANSeR algorithm, which is based on the Cork SVM algorithm that was detailed in Chapter 3, has undergone two stages of clinical trials, but final results of algorithm efficacy are not yet available (Boylan, 2015). Overcoming regulatory hurdles and demonstrating algorithm performance in varied clinical settings are challenges facing all decision support systems currently being developed in academic research. The results in this thesis show that deep learning algorithms can outperform previously developed state-of-the-art systems. One potential deployment of a neonatal seizure detection algorithm is explored in this chapter as part of portable handheld brain monitoring devices (O'Sullivan et al., 2018).

This chapter investigates different avenues for interrogating and deploying trained deep learning models:

1) It is shown how a fully convolutional network (FCN) model can be used to temporally and spatially localise and highlight short segments of EEG which the model considers to be most indicative of seizure. These segments can be further clinically analysed with regards to their origin and waveform content.
2) The hidden layer feature maps of the trained FCN models are visualised using the gradient ascent method to derive a representation of the basis for decision making. Spectral analysis is performed on these feature maps and the patterns from term and preterm models are contrasted to visualise the differences in the underlying EEG characteristics that the models learn.
3) Neonatal EEG is artificially synthesised to maximize the seizure probability at the FCN output.
4) The FCN model is efficiently integrated into an Android app to run in a low-power setting in real-time and the FCN is also integrated onto a custom embedded platform, to form part of the analytics back-end of the brain stethoscope project

7.1 Temporal and spatial localisation of salient EEG segments

One aspect of interpreting the results of a neonatal seizure detection algorithm is understanding the signal features of specific EEG windows which contributed to a seizure or non-seizure

decision. This type of analysis would give engineers and clinical experts insights into the properties of an EEG signal that are deemed to be discriminative by the algorithm. This would be an especially important analysis for reviewing false positives or false negative EEG windows. This may also be a helpful tool to utilise when an engineer is debugging an algorithm.

The fully convolutional architectures developed and analysed in Chapter 5 and Chapter 6 maintain their temporal ordering until the output of the final convolutional layer. This temporal relationship allows for the mapping of the seizure and non-seizure probabilities produced at the output of the final soft-max layer back to the original EEG inputs. This is a straight-forward method of understanding how the individual channels and the different patterns within the EEG signals affect the network output decisions.

The first piece of information which can be extracted using this temporal mapping to the multi-channel input EEG, is the location of the EEG channels where the algorithm predicts that seizure events are taking place. In cases of false positives, where the algorithm predicts a seizure but the ground truth indicates non-seizure, the channel that influenced this incorrect decision can be located. In the case of true positives, the number of channels involved during a seizure event can be quickly identified. The fully convolutional architecture is particularly beneficial in these experiments because it allows for the number of input EEG channels and the size of the EEG input window to be arbitrarily large.

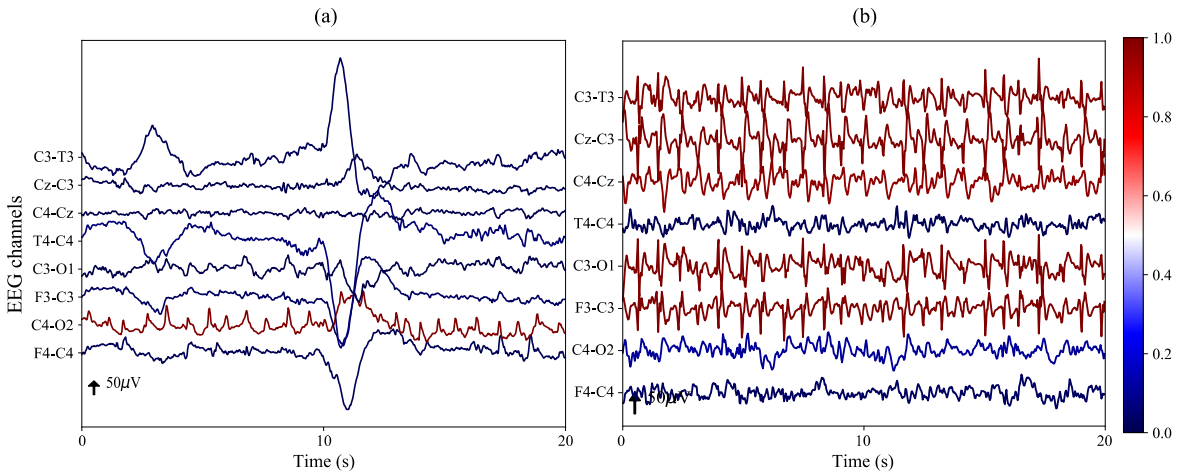


Figure 7.1 (a) and (b) show 20 second EEG segments of term neonatal seizures which have been correctly labelled as seizures by the FCN algorithm developed in Chapter 5. These examples are from two infants from Term Dataset I. These infants were not part of the training or the validation datasets for the trained networks utilised to visualise the seizure patterns. The colour-scale on the right hand side of the figure indicates the probability of seizure; red indicates that the network has assigned a high probability of seizure, blue indicates that the network has assigned a low probability of seizure.

Figure 7.1 shows a mapping of the final convolutional layer values onto the multi-channel EEG inputs. The shape of the final convolutional layer in the FCN is determined by the number of channels, the number of classes and the number of temporal samples (ordered). The probabilistic values given to each channel in Figure 7.1 are calculated by taking the average across the temporal dimension and then applying the softmax function across the values in the class dimension. This gives the seizure probability for each individual EEG channel in the input, as calculated by the deep learning algorithm during inference. Figure 7.1 (a) shows how just one bipolar channel, C4-O2, contributed to the segment being labelled as a seizure. The trained CNN gave this 20-second window a 0.983 probability of containing a seizure. Figure 7.1 (b) shows how all the channels associated with the left side of the brain and the central Cz channel are displaying patterns which the deep learning algorithm has labelled as seizure. The trained CNN gave this 20-second window a 0.997 probability of containing a seizure.

Classical machine learning methods for detecting seizures operate as sliding window classifiers; each output decision is based on a set of features which represent the full input window. The temporal ordering of feature map values in the proposed deep learning architecture allows intermediate labels that reflect the classification decision to be assigned to individual input samples within the input EEG window. The feature maps in the final convolutional layer map directly to the seizure node and non-seizure node in the output layer;

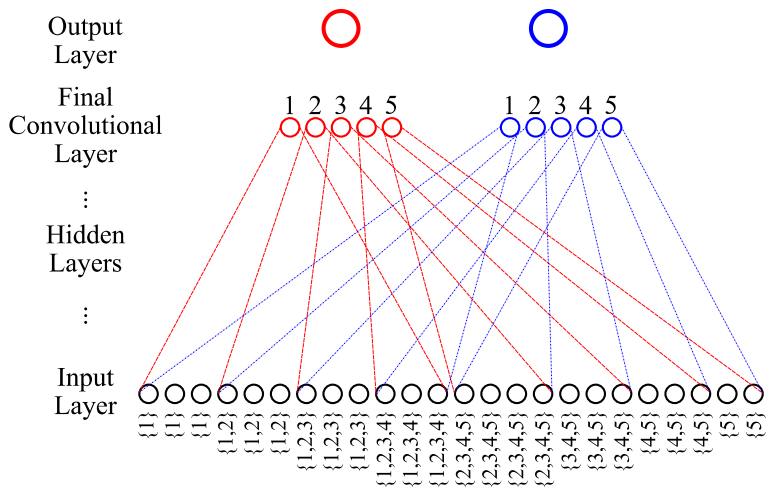


Figure 7.2 A simplistic 1-dimensional CNN is shown. It is assumed that the nodes in the final red convolutional layer map directly to the red node in the output layer, and vice versa for the blue nodes. The receptive field of each node in the final convolutional layer is shown as dashed lines of the corresponding colour. It is clear that each node in the input layer is in the receptive field of one or more of the nodes in the final convolutional layer. A set of nodes is shown beneath each input node; these represent the final layer nodes that this input sample influences via receptive field. This can be thought of as the set of final layer receptive fields that an input sample influences.

the network is trained with categorical cross-entropy where the classes are seizure and non-seizure. This forces the network to train one feature map to represent signal 'seizure-ness' and the opposite feature map to represent 'non-seizure-ness'; values in these feature maps are ordered in time.

Every value in the input EEG signal contributes to a set of values along the temporal dimension in the final convolutional layer. The mapping of receptive fields is shown for a simplistic model in Figure 7.2. The nodes that a sample in the input space contributes to are determined by the size and stride of the receptive fields for the final convolutional layer. Figure 7.2 shows how a mapping from each sample in the input space to the set of receptive fields in the final convolutional layer which that sample influences can be discovered. A seizure and a non-seizure value can be assigned to each individual sample in the input space by calculating the average value of each node in the receptive field set for that sample.

The seizure probabilities mapped to the multi-channel input in Figure 7.3 (a) are based on the mapping of the values in the seizure and non-seizure values in the final convolutional layer to each input channel, building the receptive field sets described above, and applying a point by point softmax function. The network assigned a seizure probability of 0.992 to this 20-second window. The seizure probabilities for each input sample in the sub-set of channels in Figure 7.3 (b) utilised the same receptive field mapping procedure demonstrated in Figure 7.2. Again, a point by point softmax function utilised the values in the seizure feature map and the non-seizure feature map to calculate the probability of seizure.

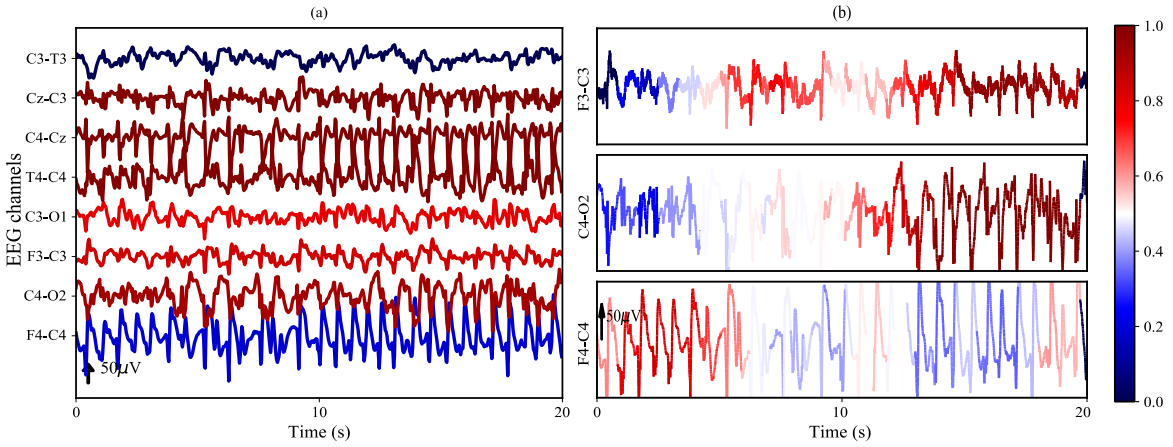


Figure 7.3 Plot (a) shows an example of a seizure event which was correctly detected by the deep learning algorithm. Plot (b) presents a more detailed view of how temporal patterns in three EEG channels contributed to the seizure probabilities in the feature maps of the final convolutional layer, based on a mapping of the input samples to their receptive fields of influence. The colour scale shows the probability of seizure.

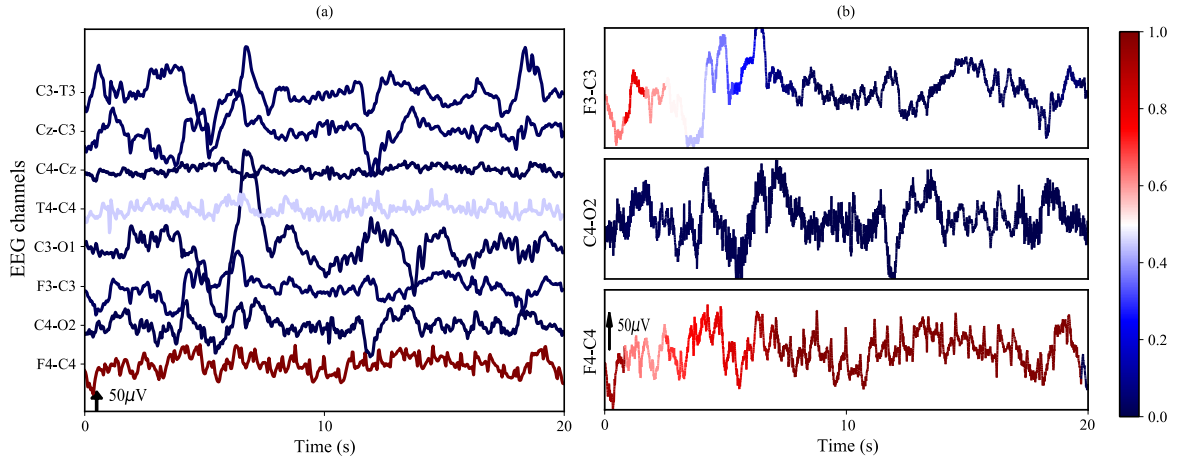


Figure 7.4 Plot (a) shows an example of a false positive error made by the deep learning algorithm. By mapping the probabilistic seizure values in the final convolutional layer to the EEG inputs in (b), the patterns which contributed to this detection error can be studied. The colour scale shows the probability of seizure.

The ability to inspect patterns in multi-channel EEG which contribute to seizure decisions could offer a data driven way of examining seizure morphologies in large datasets. This could be utilised as a clinical training tool, where an atlas of different seizure patterns is generated for a dataset of EEG from many infants. Currently the process of labelling neonatal EEG at the resolution of short individual patterns is extremely laborious; by visualising the patterns using a deep learning algorithm this process would be much faster.

Figure 7.4 shows an example of this temporal and spatial mapping of seizure decisions onto the input signals for a non-seizure segment which the network has mis-labelled as seizure. The seizure probability assigned to this window by the network was 0.737. It is clear from the mapping in Figure 7.4 (b) that the shape of the EEG in channel F4-C4 caused this window to be labelled incorrectly as a seizure event. This type of examination of false positive decisions would be useful during the algorithm research and development stage and during the algorithm clinical implementation stage. Likewise, it could be utilised during the development of artifact rejection algorithms to show the patterns which are most associated with errors. Alternatively, analysis like this could indicate to clinical staff in the NICU that there is a problem with the EEG recording setup; for example, if an electrode is causing false alarms and it needs to be reapplied.

7.2 Visualising the learned convolutional filter patterns

The most fundamental difference between previous machine learning based neonatal seizure detection systems and the algorithms developed in this thesis is the ability of deep learning

architectures to derive data-driven intermediate representations from raw input EEG. These hierarchical representations, which are the feature maps in the convolutional layers, replace the hand-engineered features that are required in machine learning based neonatal seizure detection algorithms. The end-to-end optimisation procedure ensures that the information represented in feature maps is optimised for the neonatal seizure detection task. Probing these data-driven feature map representations may give an understanding of the signal characteristics which allow the algorithm to discriminate between seizure and non-seizure EEG windows.

It may be important to reiterate at this point that in each convolutional layer a set of local filters are trained. Each of these convolutional filters generates a feature map representation based on the layer input. A feature map represents how much a particular non-linear pattern is present in the input space, these non-linear patterns are detected using the learned spatially local filters. The output of a convolutional layer represents an EEG input signal as a bank of these feature maps; each feature map relates to a different non-linear pattern.

An algorithm based on gradient ascent has been proposed for feature map visualisation in deep convolutional neural networks (Yosinski et al., 2014). This method is predicated on the understanding that the average value of a feature map, for a particular network input, is a representation of how 'activated' that feature map is. Therefore, studying the patterns which activate a feature map gives an understanding of the types of input pattern that a feature map has been trained to recognise. The aim of the gradient ascent algorithm is to engineer an input which maximises the output of a selected feature map.

First, a randomly initialised input is generated. Second, for a selected feature map in a specific hidden layer, the average value of that feature map is calculated. This forward pass of the network utilises the randomly initialised input. Third, after the forward pass, the algorithm iteratively updates the input signal to maximise this value based on the gradient. The gradient of the values in the selected feature map with respect to the network input is calculated. Steps two and three are then repeated until some stopping criteria is met. Through gradient ascent the input pattern which maximises the target feature map activation is obtained after convergence.

In this implementation of the gradient ascent algorithm $X = \{x_1, x_2, \dots, x_m\}$ is a randomly initialised m-dimensional vector. Points are sampled from a Gaussian distribution with mean of 0 and standard deviation of 50. This Gaussian distribution was chosen to represent a wide range of frequencies, and the mean and standard deviation values were selected to reflect the basic characteristics of EEG. The aim is to maximise $p_n(X)$ for the n^{th} feature map, where

$p_n(X)$ is the average of all values in that feature map for a given input X. To avoid the generation of high amplitude input values in the X vector, a regularisation term is added to $p_n(X)$. L2 regularisation is utilised as it penalises amplitude and does not result in sparse vectors. The optimisation problem becomes,

$$V(X) = p_n(X) - \lambda \|X\|_2^2 \tag{7.1}$$

$$X^* = arg \max_{X \in R^m} V(X) \tag{7.2}$$

The following gradient ascent algorithm is used to solve this optimisation problem (for the t^{th} iteration). This is iterated until convergence is obtained (or assumed), returning the input signal X^* which maximises the activation of the specified feature map. The gradient determines the direction of the change in X and η controls the gradient ascent step size.

$$X^{t+1} = X^t + \eta \frac{\partial V(X^t)}{\partial X^t} \tag{7.3}$$

Figure 7.5 illustrates the application of this gradient ascent algorithm to corresponding filters in the 5th convolutional layer of a preterm and a term network. One feature map was selected from the bank of feature maps in the 5th convolutional layer of the FCN architecture. The average of the values in the selected feature map becomes $p_n(X)$ in Equation 7.1. The gradient ascent algorithm was run for 1000 iterations, after each iteration the X vector associated with the highest $V(X)$ value was saved. After optimisation, the X^* which resulted in the highest $V(X)$ is selected.

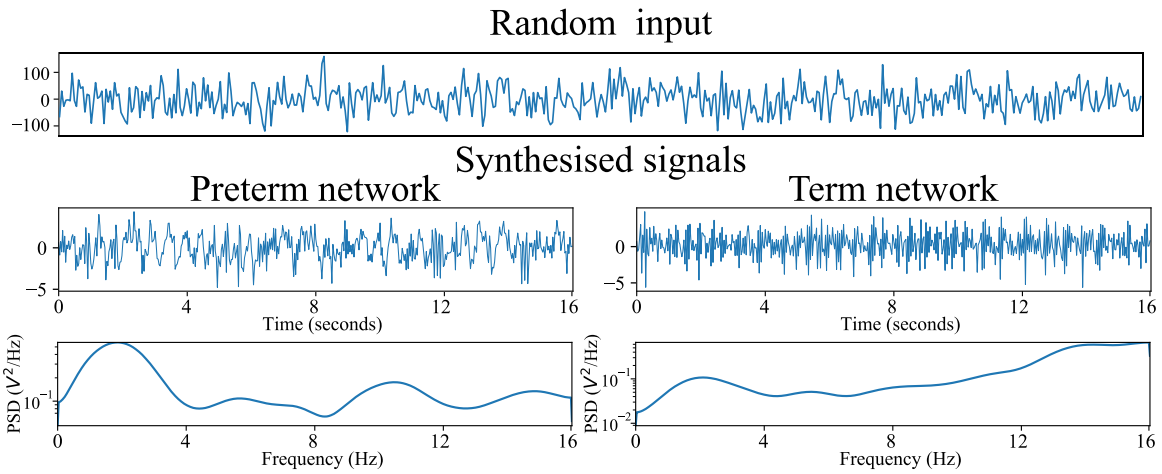


Figure 7.5 An example of signals synthesised to maximise the activation of one filter in the 5th convolutional layer of a preterm and a term network. It is important to recall that the preterm network was initialised with the term network weights through transfer learning. The same random input signal was utilised during the initial step of the gradient ascent algorithm for both networks. The PSD of these signals in the frequency domain is shown below the temporal signals. It should also be noted that the receptive field of each filter in the 5th convolutional layer of this architecture is 22 samples.

Figure 7.5 demonstrates how a randomly generated input changes after the gradient ascent algorithm has been applied; the signal was optimised with respect to a feature map in a term network and a feature map in a preterm network. In Figure 7.5, the convolutional filter which generated the feature map of the preterm network and the term network are related. The preterm convolutional filter weights were initialised with the final weights from the term network; any changes in these weights was learned during the network training procedure. The preterm and term signals, which were synthesised through gradient ascent in Figure 7.5, appear very different despite the fact that they were initialised with the same random input vector and the transfer learning of convolutional weights. The synthesised term signal appears to have higher frequency content. This difference in frequency content can be seen clearly from the power spectral density (PSD) plots of each signal. Both signals have a raised PSD centred at approximately 2Hz, but most of the power in the term signal occurs above 12Hz.

Unlike computer vision, where the increasingly complex patterns, which are hierarchically extracted and visualised from convolutional layers, can be interpreted by a human observer (Simonyan et al., 2014), analysing the neonatal EEG signals which maximally activate an individual feature map in a convolutional layer is not very informative. The gradient ascent optimisation procedure is not perfect, and the presence of local minima means that the synthesised signals can be sub-optimal. Also, the relative importance of different feature maps across a layer is not understood. Another option which may be more informative is to analyse all feature maps in a convolutional layer together. This will give a more general understanding of how signals are represented at a certain level of a deep architecture. This also allows for the comparison of different layers in the hierarchical networks.

The synthesised signals for all feature maps in a layer of a convolutional network can give an insight into the types of signals that maximise the different layers of a convolutional network. Similar investigations have been conducted for convolutional neural networks that were trained for image processing tasks. These synthesised signals are quite noisy; they do not show clear repeating patterns of activation. To display the patterns more clearly, the filters of a single layer in a single network are plotted together by converting them to the frequency domain using an fast Fourier transform (FFT) operation to calculate the PSD of the signal; the one-sided FFT of the signal was calculated using a Hanning window, the square of the absolute value of the FFT was calculated to demonstrate the power in each frequency bin. This transformation means that each feature map is represented by a 1-dimensional vector which shows the power of the signal

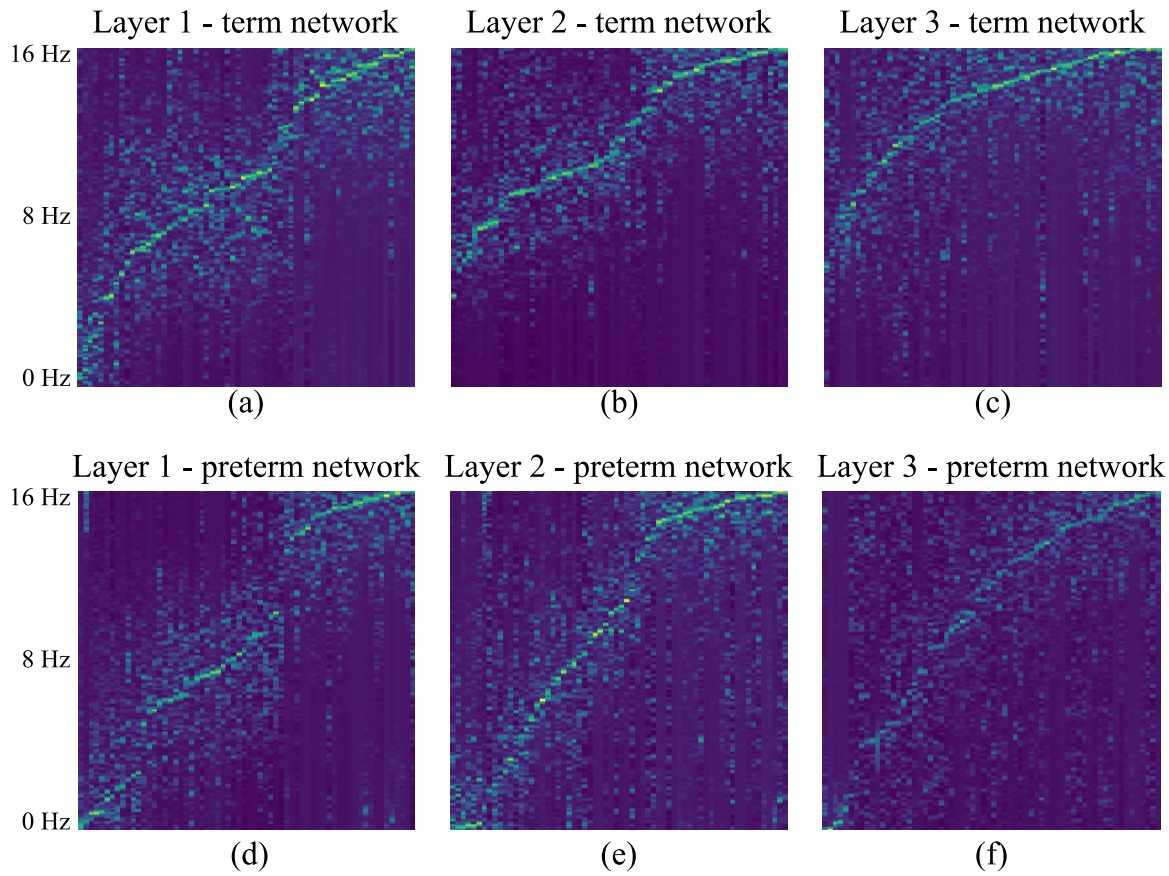


Figure 7.6 The first three convolutional layers of the term (a-c) and the preterm (d-f) FCNs. The y-axis of each plot represents different frequency bins, the x-axis shows the PSD associated with each feature map in the convolutional layer. The preterm network was optimised to detect seizures in infants in preterm group 2.

over a range of frequency bands. To plot the frequency characteristics of all feature maps in a manner that is interpretable to the human eye, they are sorted by peak frequency bin. The feature maps in a single convolutional layer are ordered based on the frequency bin of peak magnitude, from low to high frequencies. This allows the overall frequency trends of each convolutional layer to be examined, rather than attempting to interpret the noisy temporal signals for each individual filter.

Figure 7.6 shows the PSD of the signals synthesised through gradient ascent for each feature map in the first three convolutional layers of two networks. By comparing corresponding convolutional layers, the differences between the learned filters in the preterm and the term networks can be observed. Again, the preterm network was initialised with the learned weights from the term network through transfer learning, and the same signal was utilised at $t = 0$ in the gradient ascent algorithm for corresponding feature maps in the term and preterm networks. For visualisation purposes the experiments to optimise each individual feature map were

conducted twice. Each plot in Figure 7.6 consists of two PSD signals that were generated through gradient ascent for each of the 32 feature maps in a particular convolutional layer, this results in 64 individual signal PSDs, which are ordered by peak frequency, spanning the x-axis.

Due to the hierarchical nature of the convolutional neural network, each layer has a different capacity for learning signal patterns. This capacity relates to the receptive field and the convolutional depth (with non-linear activations). For example, the three sample wide convolutional filters in the first layer can learn a smaller range of input patterns than the three sample wide convolutional filters in the second layer, because the filters in the second layer exploit the nuanced representation of the signal already extracted by filters in the first layer.

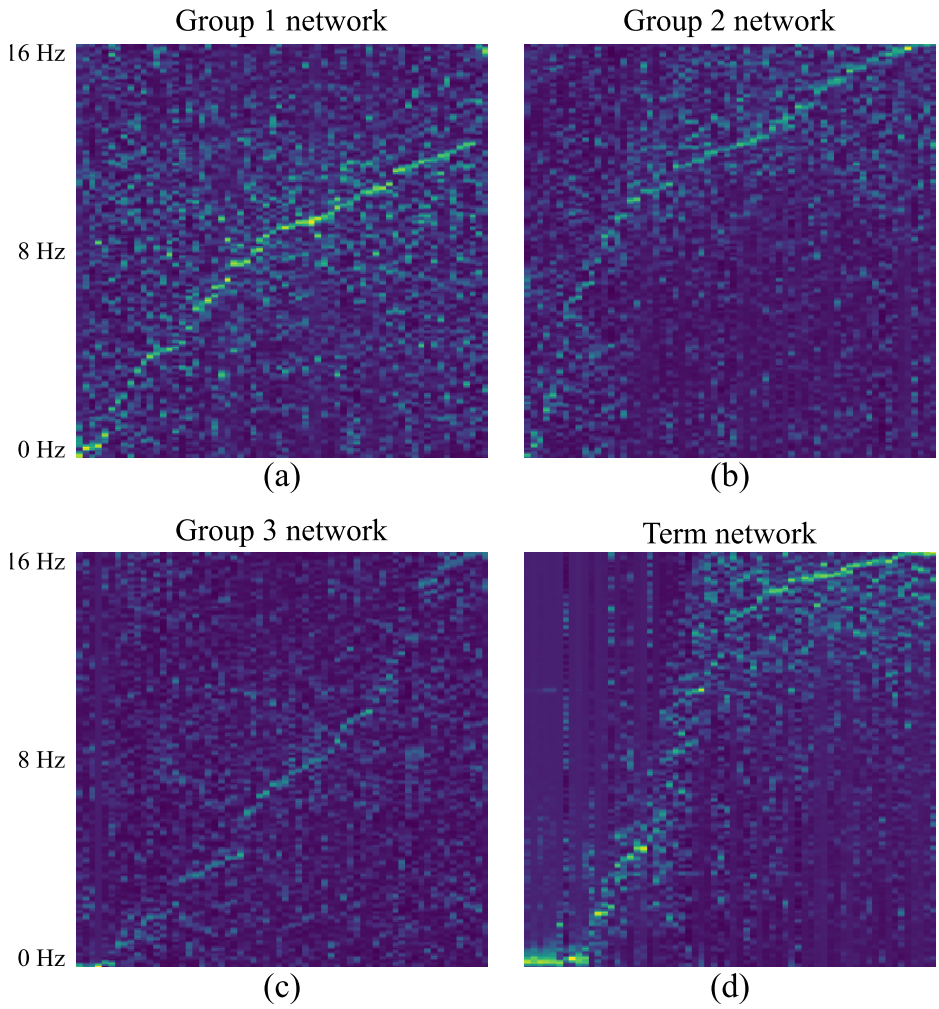


Figure 7.7 A visualisation of the 7th convolutional layer of four neonatal seizure detection networks is shown in this figure. (a), (b) and (c) show the filters from the trained preterm networks for different preterm GA groups in order of increasing age; Group 1: GA<26w; Group 2: 26w≤GA≤29w; Group 3: GA>29w. (d) shows the corresponding convolutional layer from a term network; Term: GA>38w. The y-axis of each plot represents different frequency bins, the signals associated with different feature maps in each layer are on the x-axis.

Activity in the delta band (0-4Hz) is considered to be an important feature in preterm EEG, it is associated with brain maturation (Pavlidis et al., 2017). The frequency of most preterm seizure events also lies in the delta and theta bands; making them hard to distinguish from the delta band background activity (Janáčková et al., 2016). Janáčková et al. (2016) demonstrated that the maximal frequency of seizures in infants increases with gestational age. These characteristics of neonatal seizures could account for the fact that the term filters visualised in Figure 7.6 (a-c) are more concentrated on detecting higher frequencies when compared with the preterm filters visualised in Figure 7.6 (d-f). In general, neonatal seizures (term and preterm) have been reported to have dominant frequencies ranging from 0.5-13Hz, so both networks should learn filters to extract patterns in these ranges (Kitayama et al., 2003).

In Figure 7.6 the first three convolutional layers of a term network are compared with a preterm network that was optimised to detect seizures in the gestational age (GA) group 2 infants. Figure 7.7 however contrasts three separate networks that have been optimised to detect seizures in preterm EEG, for GA groups 1, 2 and 3, with a network that has been optimised to detect seizures in term EEG. A comparison of the learned filters in a deep learning algorithm based on GA could allow clinical neurophysiologists to understand the signal characteristics required to discriminate between seizures at different developmental stages. Each filter representation in Figure 7.7 is different, but some share common characteristics. For example, Figure 7.7 (a), (b) and (d) show a number of filters that are maximally activated at low frequencies, between 0Hz and 2Hz. Figure 7.7 (c) does not have many filters that are maximally activated by signals in this frequency range. Observations like these could potentially offer insights into the characteristics of neonatal seizures and how they vary with GA.

7.3 Synthesising artificial neonatal EEG

Another aspect of the data-driven filters in a deep learning algorithm which can be exploited is the potential for EEG signal generation. The most commonly employed deep learning synthesis algorithms are Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), but the gradient ascent algorithm can also be utilised to exploit the characteristics of the already trained deep learning classification algorithms to generate artificial EEG examples. This could be a useful alternative to using real EEG to ensure complete data anonymisation during public presentations or to augment clinical training data.

The gradient ascent algorithm can be utilised to synthesise signals that maximally activate either the seizure or the non-seizure feature map in the final convolutional layer of the trained

FCNs. The resultant engineered artificial signals represent what the deep learning model considers to be *ideal* seizure or non-seizure patterns.

In this method, the previously trained deep learning models are utilised. The same gradient ascent algorithm is employed, as shown in Section 7.2. Previously it has been assumed that the values of the individual feature maps in each layer are not connected, and that they can be optimised independently. Due to the softmax non-linearity at the network output, the values of the feature maps in the final convolutional layer are not independent of one another. The softmax function generates a probabilistic distribution from a vector based on the differences between values. The probability attributed to each value is related to the ratio of the exponent of that value relative to the sum of the exponents of all values in the vector.

To synthesise EEG signals the loss needs to reflect both feature maps in the final convolutional layer. p_n is the mean of all values in the n^{th} feature map of the final convolutional layer; $n = 0$ represents the non-seizure feature map and $n = 1$ represents the seizure feature map in the trained networks. To optimise the network to synthesise an example seizure signal ($X_{seizure}$), the following optimisation problem is solved

$$V_{seizure}(X) = \frac{e^{p_1(X)}}{e^{p_0(X)} + e^{p_1(X)}} = \frac{1}{e^{p_0(X) - p_1(X)} + 1}. \tag{7.4}$$

$$X_{seizure} = arg \max_{X \in R^m} V_{seizure}(X) \tag{7.5}$$

This is equivalent to maximising $(p_1(X) - p_0(X))$ as the target for seizure signal generation. To generate non-seizure signals the target is to maximise $(p_0(X) - p_1(X))$. As described in Section 7.2, L2 regularisation is used to avoid generating high magnitude signals.

Figure 7.8 shows an example of a synthesised term non-seizure signal and a term seizure signal. The seizure signal shows repetitive spikes and appears to have more low frequency content than the non-seizure signal. The network predicts that the synthesised non-seizure signal has a probability of 0.00006 of being a seizure window. The network predicts that the synthesised seizure signal has a probability of 1.0 of being a seizure window. These scores imply that the network has generated signals that are representative of what it considered to be non-seizure and seizure ideal examples. This does not mean that these signals are representative of realistic EEG, but this exercise may give engineers and medical practitioners insights into what kinds of signal patterns the networks are learning to detect.

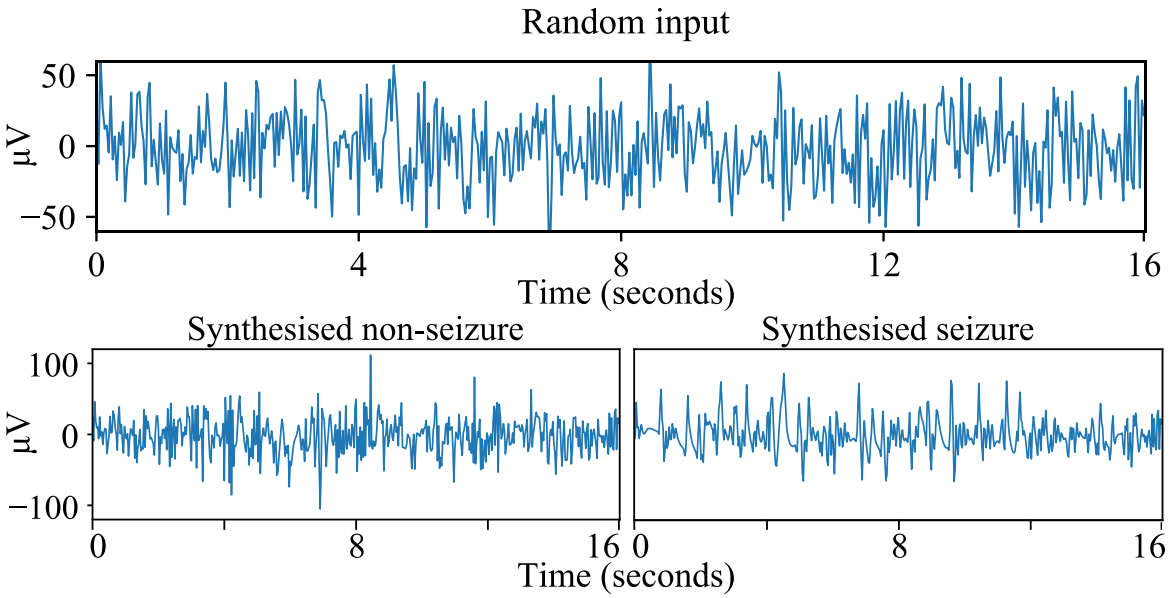


Figure 7.8 Synthesised term EEG signals which were generated through gradient ascent. A network optimised to detect seizures in term EEG was utilised to synthesise these signals. A signal to maximise the non-seizure probability and a signal to maximise the seizure probability were both generated from the same initial random input signal.

Figure 7.9 shows an example of a synthesised preterm non-seizure signal and a synthesised preterm seizure signal. The network employed to synthesise these signals was optimised to detect seizures in group 1 preterm infants; group 1 infants are those of less than 26 weeks GA. These extremely preterm infants are known to exhibit large inter-burst intervals during non-seizure EEG. In Figure 7.9, the burst of high amplitude signal at the 8-second mark and again at the 12-second mark in the non-seizure EEG signal could potentially be interpreted as bursts, with the low amplitude period separating these bursts being interpreted as the inter-burst interval. It is known that bursts and inter-burst intervals are important features in preterm EEG that relate to GA and normal development (Pavlidis et al., 2017). The network predicts that the synthesised non-seizure signal has 0.004 probability of being a seizure window. The network predicts that the synthesised seizure signal has 0.948 probability of being a seizure window. Even though these signals have been optimised for classification by this algorithm, it is not trivial to understand the signal features which cause the network to have such certainty about the classification decisions.

An alternative method of physiological signal generation is to utilise a GAN. GANs are capable of learning realistic data generators through deep learning (Goodfellow et al., 2014). This learning takes place as part of a contest where a generative network and discriminative network are optimised as part of a minimax game. The generative network aims to generate high quality

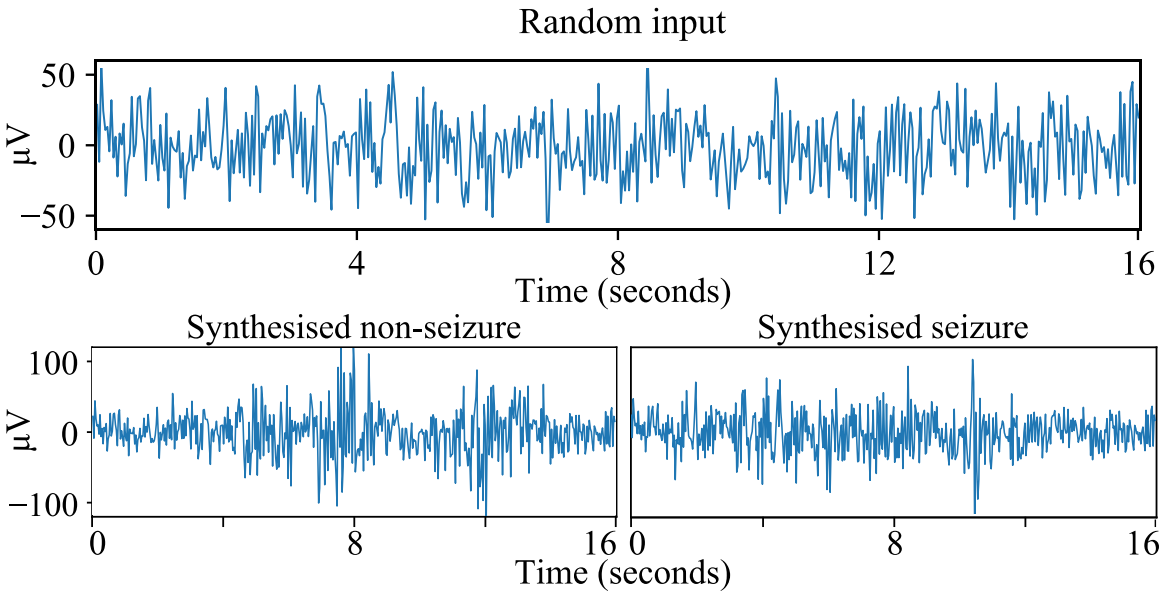


Figure 7.9 Synthesised preterm EEG signals which were generated through gradient ascent. A network optimised to detect seizures in preterm EEG was utilised to synthesise these signals. The preterm network was initialised with weights from a trained term network though transfer learning, the network was then further trained to detect seizures in preterm EEG.

synthetic samples which will cause the discriminative network to fail. The discriminative network aims to determine if the sample is real or if it is a fake that was synthesised by the generative network. This process has been utilised to generate electrocardiogram (ECG) data (Delaney et al., 2019), and could further be employed to generate realistic EEG signals.

7.4 Algorithm deployment on a low-cost portable device

The translation of engineering research to clinical settings is an important aspect to consider when developing new methods of interpreting physiological signals. In this subsection, the computational requirements of the developed models are shown in the context of Android app implementation and a custom embedded systems implementation using an analogue front-end application specific integrated circuit (ASIC) and a low-power microcontroller unit. Light-weight versions of the models developed in Chapter 5 are created to assess the trade-off between classification performance and computational cost.

An advantage of the neonatal seizure detection solutions that have been designed, developed, and tested in this work is that they have a relatively simple implementation at the inference stage, and they consist of a relatively small number of learned parameters which need to be stored in memory. The Cork SVM algorithm has a feature set of 55 features and utilises approximately 10,000 support vectors for classification, this amounts to over 550,000 parameters which need to be stored in memory to implement this algorithm (O'Shea et al.,

2017; Temko et al., 2011). The presence of a feature extraction also substantially increases the time and computational power required to implement machine learning algorithms. The SVM based algorithms require a feature extraction stage, where windows of input EEG must be transformed into features before the classification algorithm stage; feature extraction results in additional computational complexity and significantly adds to the algorithm run time. For example, with its computationally hungry time-frequency and correlation based features, the Helsinki SVM algorithm when tested on the EEG from Term Dataset I required over 4.5 hours to process each hour of raw EEG on a PC with an Intel i5 processor and 8GB of RAM (O'Shea et al., 2020).

In contrast to the parameter heavy and computationally intensive machine learning algorithms, a trained FCN relies only on spatially local matrix multiplication, non-linearity functions, and pooling operations at inference time. The fully convolutional architectures, which are a feature of this work, have fewer parameters than similar convolutional neural networks, which include parameter heavy densely connected layers. The ten layer FCNs utilised in Chapter 6 have approximately 25,500 parameters, 20 times less than the Cork SVM algorithm. Additionally, these ten layer FCNs can process one hour of raw temporal EEG in less than one minute when run on a PC (Intel i5 processor and 8GB of RAM, same as above) using only the CPU for computation.

The brain stethoscope project aims to develop a portable device which facilitates faster and more accessible neonatal EEG monitoring and includes artificial intelligence powered decision support. The device is designed to be low-cost, battery powered, and require minimal setup time; this is in stark contrast to current cot-side neonatal monitoring devices which cost thousands of euros, require significant expertise and time to setup, and are typically housed on trolleys due to their large size (O'Sullivan et al., 2018). The investment and expertise required to purchase and utilise these traditional EEG monitoring systems is usually only available in tertiary care medical facilities (Lloyd et al., 2015). Realisation of the brain stethoscope project would facilitate neonatal brain monitoring in under resourced settings; this project has been in development since 2015.

An initial proof of concept of the brain stethoscope project saw the implementation of EEG acquisition, visualisation, and decision support as part of an Android application to be run on an off-the-shelf Android compatible device. A second bespoke implementation of the brain

Figure 7.10 Diagram of the App-based variant of the brain stethoscope which consists of portable EEG acquisition, communication and Android app-based interpretation framework. The deep learning algorithms were incorporated into the analytics back-end of the device, which was hosted on a tablet. The Android application was built to visualise the EEG signal and a traffic-light indicator showed the presence of neonatal seizures based on the probabilistic output from a trained model.

stethoscope project saw the development of a custom printed circuit board which enables acquisition and processing (including deep learning inference) of EEG signals. Fully convolutional deep learning algorithms which were developed over the course of this thesis project have been implemented to facilitate on the edge decision support in both versions of the brain stethoscope project.

As part of a wider project to develop a portable neonatal EEG evaluation and decision support Android application, two fully convolutional architectures were realised in Java and their performance on a commercially available Android compatible tablet device was assessed (O'Sullivan et al., 2018). The Android application was developed and tested on a Samsung A6 tablet; it was designed to receive data from a single channel of EEG via Bluetooth. The conceptual diagram of the developed device is shown in Figure 7.10. The EEG acquisition hardware interacts with the interpretation software resulting in a real-time assessment of EEG based on a traffic light indicator.

Table 7.1 A comparison of performance results and the associated computational cost of two FCNs when implemented on a mobile platform. The performance results reported are the result of LOO testing on the Cork term dataset. The open source Battery Historian program was utilised to measure the RAM, power consumption and CPU usage (O'Sullivan et al., 2018).

Network	AUC	AUC 90	GDR (0.5FD/h)	# params	CPU (%)	RAM (MB)	Power (mAh)
Six layer	97.1%	82.9%	78.8%	17,058	21.29±2.8	23.33±1.18	30.3
Eleven layer	97.7%	86.5%	83.2%	28,642	25.63±1.0	25.96±0.72	38.21

As a proof of concept, the application was designed to accommodate the processing of a single channel EEG in real-time. An eleven layer FCN model implemented on this device was trained on strong EEG labels and utilised only one dimensional convolutional and pooling filters (O'Shea et al., 2018, 2017). A light-weight version of the model was also implemented which consisted of six convolutional layers with smaller resultant receptive field. As it can be seen Table 7.1, the increased depth and receptive field in the eleven layer network lead to improved classifier performance, but it also resulted in increased computational and memory cost (O'Shea et al., 2018; O'Sullivan et al., 2018).

Implementing the EEG recording, processing, and interpretation software on an off the shelf Android device validated the system design and acted as an initial proof of concept for the brain stethoscope project. The success of this Android implementation motivated further optimisation of the design, implementation, and testing of a custom embedded platform for portable neonatal EEG analysis. The deep learning neonatal seizure detection algorithms integrated into the device were based on the architectures designed and validated as part of this thesis.

A custom embedded system platform was designed for high accuracy EEG acquisition and analysis. Figure 7.11 illustrates how the different components of this system interact together to provide EEG acquisition, processing, visualisation, and decision support. The platform comprises of power management, analogue-to-digital conversion, and micro-controller circuitry. A FCN term deep learning seizure detection algorithm was implemented on the micro-controller unit.

Implementation on a low-cost and battery-powered platform imposed memory and power limitations. The first major limitation, that impacted the deep learning architecture integration

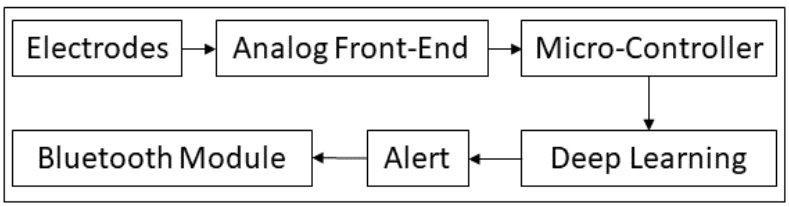


Figure 7.11 Block diagram of the on-the-edge implementation of the brain stethoscope project. Electrodes record the EEG signal. The signal is digitised and filtered in the analogue front-end circuitry. The EEG is processed in the micro-controller unit and here deep learning is utilised to generate a seizure or non-seizure decision. Based on the deep learning algorithm output an alert may be given to provide real-time clinical decision support. Finally the EEG signal and the ouput of the AI algorithm are wirelessly transmitted using the Bluetooth module for visualisation, further monitoring, or for long-term storage.

Table 7.2 The 10 layer fully-convolutional architecture with increased pool stride.

Layer	Filter width	Stride	Non-linearity	Receptive field
Convolution	3	1	ReLU	3
Convolution	3	1	ReLU	5
Convolution	3	1	ReLU	7
Batch norm				7
Average pool	4	4		10
Convolution	3	1	ReLU	18
Convolution	3	1	ReLU	26
Convolution	3	1	ReLU	34
Batch norm				34
Average pool	4	4		46
Convolution	3	1	ReLU	78
Convolution	3	1	ReLU	110
Convolution	3	1	ReLU	142
Batch norm				142
Average pool	4	4		190
Convolution	3	1	ReLU	**318**
Average pool	Global			400
Softmax				400

into the embedded system, was the data sampling rate of the on-board analogue-to-digital circuitry. The system samples EEG at 250Hz; the networks developed in previous chapters of this work were optimised to detect seizures in EEG which was recorded at 256Hz and downsampled to 32Hz. The memory limitations of the system prevented resampling from 250Hz to 32Hz. Instead, all networks were retrained on EEG which was resampled at 50Hz, the architecture was adapted slightly to account for this change in input EEG representation.

Previous work has shown the importance of receptive field on the performance of FCN algorithms for neonatal seizure detection (O'Shea et al., 2020). Changing the input representation from 32Hz to 50Hz sampling rate implies that the number of samples in the receptive field should be increased to ensure that filters in the final convolutional layer have an equivalently wide operating range when measured in seconds of EEG. In the ten layer FCN architecture developed in Chapter 5, the receptive field of the final layer was 172 samples, this is equivalent to 5.4 seconds at an input sampling rate of 32Hz. An equivalent architecture was developed and trained with 50Hz EEG signals with the aim of implementing it on this device. To increase the receptive field in the final convolutional layer, the stride of each pooling layer was increased from three samples to three samples. Table 7.2 shows the network architecture with increased pool stride and the effects that this has on receptive field. The receptive field in

Table 7.3 A comparison of two fully convolutional algorithms with 10 convolutional layers. The number of parameters in each layer and the approximate number of floating point operations (FLOP) required to process an 8-second window of EEG sampled at 50Hz input are presented. Pooling, normalisation, and non-linearity layers are not shown in this analysis.

Convolutional layer index	32 feature map model		16 feature map model	
	Parameters	Operations (kFLOP)	Parameters	Operations (kFLOP)
1	64	19.1	128	38.2
2	784	304.1	3104	1216.5
3	784	302.6	3104	1210.4
4	784	75.3	3104	301.1
5	784	73.7	3104	294.9
6	784	70.7	3104	288.8
7	784	16.1	3104	64.5
8	784	14.6	3104	58.4
9	784	13.1	3104	52.2
10	98	0.2	194	0.4

the final convolutional layer is 318 samples; this equates to 6.4 seconds at an input sampling rate of 50Hz.

The second major limitation was the large number of network parameters in previously developed term fully convolutional networks (see Chapter 5). The computational requirements of this architecture were too memory intensive for the embedded system. To overcome this limitation the number of feature maps in each convolutional layer was reduced, this led to a reduction in the flash storage and the random access memory required to run the deep learning

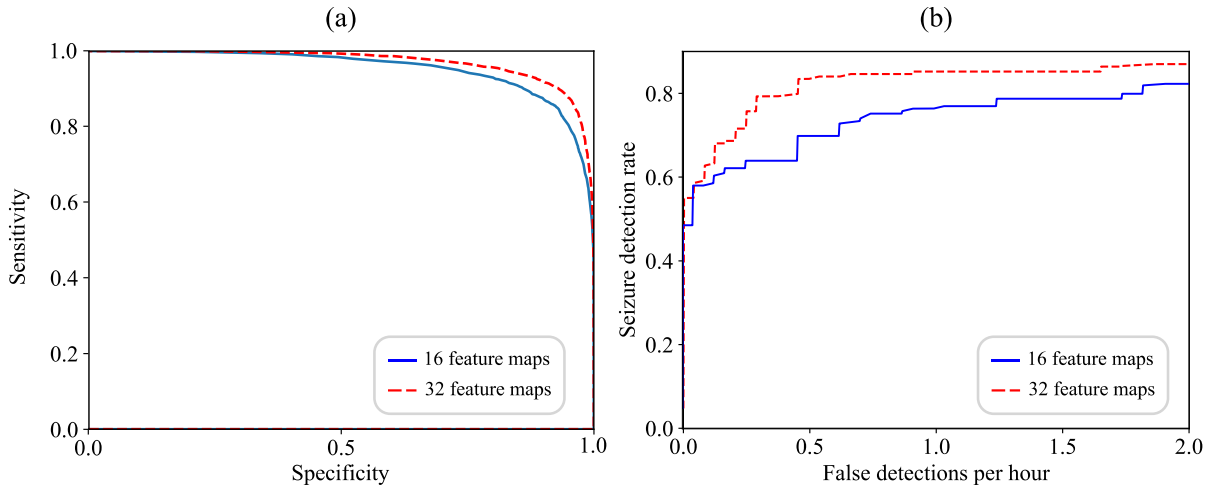


Figure 7.12 A comparison of network performances with a varying number of feature maps when tested on a single infant. (a) presents the ROC curves for the 16 feature map and 32 feature map models. (b) presents the variation in the seizure detection rate for a range of thresholds that result in a certain number of false detections per hour.

algorithm. Table 7.3 shows how the reduction in the number of feature maps in each layer impacts the layer parameters and the computational requirements. The FLOP requirement of each convolutional layer is calculated by multiplying the input channel dimension, the output channel dimension, the size of the feature map filters and the number of operations together for each layer. The reduction of feature maps and subsequent reduction in computational requirements leads to a reduction in algorithm performance. Figure 7.12 presents the epoch based and event-based sensitivity of the algorithm over a range of operating points, assessed on a 24h recording of neonatal EEG from one infant. Reducing the number of feature maps in each convolutional layer does come at the cost of decreased accuracy, but this trade-off is required to meet the computational restrictions imposed by the embedded device.

7.5 Discussion

This chapter has focused on exploiting the properties of trained deep learning algorithms to derive engineering insights and clinical insights from the data-driven feature extraction layers and to translate this research into real-time portable clinical decision support for use in low resource settings. The use of a fully convolutional architecture throughout this body of work has influenced the nature of model analysis and also the suitability of these algorithms for embedded systems integration.

In this chapter, model characteristics were analysed to generate insights which are potentially clinically relevant. Investigating deep learning model properties is not straightforward and the methods utilised in this chapter were not analysed with objective metrics, rather visual representations of learned weights and of signals, which maximally activate different parts of the models, were represented in plots and the apparent differences between preterm and term characteristics were discussed. The term EEG trained models developed in Chapter 5 and the preterm EEG trained models developed in Chapter 6 were utilised to show contrasts between the patterns learned to facilitate seizure detection in term and preterm settings. The clinical significance of patterns and filter characteristics were not explored, this type of analysis would require a large amount of clinical insight and was considered to be outside the scope of this chapter.

Insights into the functionality of the developed models can be obtained from the trained networks. The architectural choices that were made during this work influenced the types of network insights which could be derived and explored. The architectural characteristics of the fully convolutional models allowed patterns which contributed to seizure decisions to be

localised in time and space. If fully connected layers had been employed in the models or if the convolutional weights had been extended to have inter-channel dimensionality then this mapping would not have been as straightforward.

Deep learning algorithms have proven successful in a wide variety of classification tasks, but understanding their inner workings remains challenging. There are many methods of probing trained deep networks in image processing, but most of these rely on a human's intuitive understanding of images (Zeiler and Fergus, 2013). Frequency domain analysis offers a method of understanding and visualising the hidden layers of networks which are designed to classify signals (Lee et al., 2018). This chapter has shown the contrasts between the underlying signal patterns which the different networks have been tuned to detect across different preterm and term networks. The differences in these hidden feature extraction layers reinforces the results from Chapter 6 which showed that the features developed to detect seizures in term EEG are not optimised to detect seizures in preterm EEG and it reinforces the understanding that different networks are required to effectively detect seizures in the EEG of different GA groups.

The exploration of EEG signal synthesis in this work contributes to the understanding of the patterns which influence trained deep learning algorithmic decisions and it explores the possibility of generating synthesised EEG signals to augment clinical and machine learning research. Understanding the patterns which contribute to an algorithm's decisions is important to give engineers and clinical practitioners insights into important signal characteristics. In the case of algorithm mis-classifications the limitations of the trained algorithm can be explored. Deeper understanding of false negative decisions (missed seizure events) may highlight seizure morphologies that are not present in the training dataset.

Labelled neonatal EEG signals are difficult to obtain due to environmental constraints in the NICU, privacy and ethical concerns, and the time and expertise associated with the generation of accurate labels. The generation of realistic and varied EEG data could be used in clinical training, where fully anonymised data are required, or it could be utilised to augment the training datasets for deep learning or machine learning algorithms. The signal synthesis routine demonstrated in this chapter would need to be further developed and validated, but it does demonstrate that trained deep networks can be employed for surrogate data generation.

Deep learning algorithms have a reputation for requiring intensive computational resources, which would imply that they are not practical for implementation in environments with limited power and computational resources. This mis-understanding stems from the distinction

between the training and inference stages of deep learning algorithm development. In Chapter 4, the hardware requirements for deep learning algorithm training were outlined; it was highlighted that most deep learning algorithms are trained using specialised processing units, usually GPUs or TPUs, due to the quantity of mathematical operations performed during training, the iterative nature of model optimisation, and the high dimensional input spaces. In contrast to the intensive training stage, the inference stage of most deep learning algorithms does not require parallel processing or specialised hardware; this does however depend on the complexity of the model and the required velocity of the data being processed during deployment. The Android implementation and the custom edge device implementation of neonatal seizure detection algorithms in this chapter show that fully convolutional algorithms are suitable for deployment on mobile, low-power platforms to give real-time seizure decision support. This finding shows the potential for deep learning based algorithms to contribute to decision support in clinical settings and also to have an impact in low-resource settings outside the NICU; for example in the developing world or in local neonatal units outside major neonatal hospitals where access to bulky and costly traditional EEG monitoring devices is not available.

The translation of trained deep learning algorithms onto portable monitoring devices has proven the practical benefits of the developed deep learning seizure detection algorithms. The developed algorithms are less computationally heavy and memory intensive when compared to previous SVM based algorithms due to the removal of any feature extraction stage and the relative simplicity of the inference stage implementation. The Cork SVM algorithm relies on 20 times more parameters when compared with the developed algorithms and the Helsinki SVM algorithm takes 300 times longer to process EEG than the developed algorithms when tested under the same computational conditions. These numbers indicate that the computational and memory burden associated with deep learning is substantially less than that of the machine learning baseline neonatal seizure detection algorithms. The implementation of the deep learning algorithms on resource constrained devices did come at the cost of a reduction in performance metrics. However, the potential for real-time, on the edge, decisions would provide the potential for brain monitoring earlier and in more clinical locations; this is not considered to be a replacement for long-term EEG monitoring by clinical neurophysiology experts.

7.6 Conclusion

The analysis completed in this chapter demonstrates the potential for clinical insights to be derived from trained deep networks. Deep learning insights can provide feedback to medical

practitioners and machine learning engineers and the examples in this chapter show the feasibility of the development of algorithm interrogation frameworks. Further work and input from clinical practitioners are required to validate the signals generated through the gradient ascent algorithm. This validation procedure should verify that the signal properties reflect that of realistic EEG.

Clinical translation is an important aspect of biomedical research. The experiments in this chapter have indicated that the developed deep learning architecture is suitable for resource constrained device implementation. This is in part due to the lightweight architecture resulting from the fully convolutional properties and also due to the removal of any feature extraction requirements. The implementation comes at the cost of performance as the preprocessing, network architecture and postprocessing were all adapted to fit on the embedded system platform.

7.7 References

Andreotti, F., Phan, H., & De Vos, M. (2018). Visualising Convolutional Neural Network Decisions in Automatic Sleep Scoring. *CEUR Workshop Proceedings*, 70–81.

Boylan, G. (2015). *ANSeR- The Algorithm for Neonatal Seizure Recognition Study*. https://clinicaltrials.gov/ct2/show/NCT02160171

Choo, J., & Liu, S. (2018). Visual Analytics for Explainable Deep Learning. *IEEE Computer Graphics and Applications*, *38*(4), 84–92.

Delaney, A. M., Brophy, E., & Ward, T. E. (2019). Synthesis of Realistic ECG using Generative Adversarial Networks. *ArXiv:1909.09150 [Cs]*.

Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., & Acharya, U. R. (2018). Deep learning for healthcare applications based on physiological signals: A review. *Computer Methods and Programs in Biomedicine*, *161*, 1–13.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems 27*, 2672–2680.

Janáčková, S., Boyd, S., Yozawitz, E., Tsuchida, T., Lamblin, M.-D., Gueden, S., & Pressler, R. (2016). Electroencephalographic characteristics of epileptic seizures in preterm neonates. *Clinical Neurophysiology*, *127*(8), 2721–2727.

Kitayama, M., Otsubo, H., Parvez, S., Lodha, A., Ying, E., Parvez, B., Ishii, R., Mizuno-Matsumoto, Y., Zoroofi, R. A., & Snead, O. C. (2003). Wavelet analysis for neonatal electroencephalographic seizures. *Pediatric Neurology*, *29*(4), 326–333.

Lee, J., Park, J., Kim, K. L., & Nam, J. (2018). SampleCNN: End-to-End Deep Convolutional Neural Networks Using Very Small Filters for Music Classification. *Applied Sciences*, *8*(1), 150.

Lloyd, R. O., Goulding, R. M., Filan, P. M., & Boylan, G. B. (2015). Overcoming the practical challenges of electroencephalography for very preterm infants in the neonatal intensive care unit. *Acta Paediatrica*, *104*(2), 152–157.

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

O'Shea, A., Lightbody, G., Boylan, G., & Temko, A. (2020). Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture. *Neural Networks*, *123*, 12–25.

O'Shea, A., Lightbody, G., Boylan, G., & Temko, A. (2018). Investigating the Impact of CNN Depth on Neonatal Seizure Detection Performance. *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 5862–5865.

O'Shea, A., Lightbody, G., Boylan, G., & Temko, A. (2017). Neonatal seizure detection using convolutional neural networks. *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6.

O'Sullivan, M., Gomez, S., O'Shea, A., Salgado, E., Huillca, K., Mathieson, S., Boylan, G., Popovici, E., & Temko, A. (2018). Neonatal EEG Interpretation and Decision Support Framework for Mobile Platforms. *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 4881–4884.

Pavlidis, E., Lloyd, R. O., Mathieson, S., & Boylan, G. B. (2017). A review of important electroencephalogram features for the assessment of brain maturation in premature infants. *Acta Paediatrica (Oslo, Norway: 1992)*, *106*(9), 1394–1408.

Ramantani, G., Schmitt, B., Plecko, B., Pressler, R. M., Wohlrab, G., Klebermass-Schrehof, K., Hagmann, C., Pisani, F., & Boylan, G. B. (2019). Neonatal Seizures—Are We there Yet? *Neuropediatrics*, *50*(5), 280–293.

Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical Decision Support in the Era of Artificial Intelligence. *JAMA*, *320*(21), 2199–2200.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *ArXiv:1312.6034 [Cs]*.

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv:1409.1556 [Cs]*.

Temko, A., & Lightbody, G. (2016). Detecting Neonatal Seizures With Computer Algorithms. *Journal of Clinical Neurophysiology*, *33*(5), 394–402.

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

Vanhatalo, S. (2011). Development of neonatal seizure detectors: An elusive target and stretching measuring tapes. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *122*(3), 435–437.

Vilamala, A., Madsen, K. H., & Hansen, L. K. (2017). Deep convolutional neural networks for interpretable analysis of EEG sleep stage scoring. *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6.

Yosinski, J., Clune, J., Fuchs, T., & Lipson, H. (2014). Understanding neural networks through deep visualization. *In ICML Workshop on Deep Learning*.

Zeiler, M. D., & Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. *ArXiv:1311.2901 [Cs]*.

Chapter 8: Conclusions

8.1 Concluding summary and novel contributions

This thesis detailed the development and validation of automated seizure detection algorithms using deep learning methods with the aim of providing accurate clinical decision support in the neonatal intensive care unit (NICU). Deep learning provided a novel method of learning data-driven, discriminative features from raw temporal inputs through end-to-end algorithm optimisation. Innovative architectures and adapted algorithm training frameworks were developed to efficiently utilise the available data. The potential of the developed deep learning architectures was demonstrated in both term and preterm infants, the ability to analyse trained networks was explored, and avenues for clinical translation through mobile embedded systems were presented.

A diverse range of topics have been examined in this thesis, all relating to the development of deep learning algorithms for neonatal seizure detection. More specifically, methods of developing neonatal seizure detection algorithms which do not rely on hand-crafted features and achieve state-of-the-art performances have been proposed, developed, and validated. This section summarises the conclusions of each chapter in terms of area of focus, technical challenges, and experimental findings.

Chapter 2 outlined relevant clinical information regarding neonatal brain development and monitoring, with a special focus on EEG monitoring. This chapter also provided further information about neonatal seizures, the importance of detecting seizure events, and the challenges associated with identifying seizure characteristics in the vulnerable population of neonates in the NICU. The practicalities of developing automated methods for detecting seizures in neonatal EEG were outlined, and the datasets used to train and test the algorithms developed in this work were detailed.

Chapter 3 presented the methodologies which underpin the experimental work in this thesis. A selection of neonatal seizure detection algorithm frameworks, which have been developed over the past 30 years, were examined. The clinical approach for detecting neonatal seizures in EEG signals was outlined, methods based on heuristic rules and thresholds were described and two data-driven machine learning algorithms were explained in more detail. These two machine learning algorithms represent state-of-the-art, publicly available neonatal seizure

detection algorithms and were employed as comparative baselines throughout this work. This chapter also included technical information about EEG signal preprocessing and postprocessing and important concepts regarding machine learning algorithm development, such as performance assessment. Importantly, Chapter 3 described in detail the features of deep learning networks and algorithms that were utilised for deep learning model development in this work.

In **Chapter 4,** an initial exploration of the suitability of deep learning for the neonatal seizure detection application was outlined. A variety of different neonatal EEG representations which could be utilised in conjunction with a deep learning algorithm were presented. Many works have applied deep learning to EEG signals, but very few publications have proposed an end-to-end deep learning solution for neonatal seizure detection - therefore even the initial experiments exploring this space represented novel research contributions (Craik et al., 2019; Roy et al., 2019). Empirical results of preliminary experiments were used to select an appropriate EEG representation to replace the feature set representation which had been utilised in previous machine learning based algorithms. The results in this chapter did not approach state-of-the-art performances, but they did show that deep learning algorithms are capable of learning internal hierarchical representations which can be applied to detect seizures in neonatal EEG. The experimental results in this chapter motivated further development of deep learning neonatal seizure detection algorithms.

A framework for developing a neonatal seizure detection algorithm for term EEG through end-to-end optimisation of feature extraction and classification was developed in **Chapter 5**. The chosen deep learning algorithm was a fully convolutional neural network (FCN) which operates on raw EEG signals; the use of this architecture is novel in the area of neonatal physiological signal processing. First, it was shown that, given the same initial conditions, the 1D FCN algorithm, which learns a representation from large amounts of single channel raw annotated data, outperforms an SVM-based system, which uses a set of hand-crafted features. Second, it was shown that the width of the receptive field in the final convolutional layer has the biggest influence on network performance when compared to the network depth or overall number of tuneable parameters. Finally, it was shown that the designed novel 2D FCN architecture achieved superior performance by exploiting much larger amounts of data with weak labels. Many existing machine learning based neonatal seizure detection algorithms rely on strongly labelled datasets, which greatly reduces the available training data (Tapani et al., 2019; Temko et al., 2011). This work leveraged the existing feature based neonatal seizure

detection algorithms for comparative purposes; this comparative framework ensured a clear understanding of the ability of deep learning based algorithms to outperform machine learning based algorithms. By designing an algorithm which can learn from weakly labelled data, this work provides a path to accelerate the development of algorithms in the area of neonatal EEG by circumventing the need for laborious precise per-channel labelling of datasets. The results in this chapter indicated that deep learning algorithms could outperform feature based machine learning algorithms in detecting seizures in term EEG without the requirement for bespoke feature engineering. The developed algorithms were also faster and more parameter efficient than the machine learning baselines.

Chapter 6 explored the development of a seizure detection algorithm for preterm EEG; this represents the first time that an automated seizure detection system has been developed specifically for preterm infants. This work contributed three important messages. First, initial experiments showed that the performance of algorithms developed for term EEG did not achieve the same performance levels when assessed on preterm EEG. Second, simply retraining algorithms designed for term EEG with a dataset of preterm EEG did not substantially increase the performance of either a feature-based machine learning algorithm or a data-driven deep learning algorithm. Finally, a novel end-to-end deep learning approach was proposed; the developed algorithm utilised three gestational age (GA) specific models and overcame limited data availability by weighting data based on GA group membership, utilising transfer learning from the term model, and learning from the more readily available weak seizure labels in the training stage. The results obtained using deep learning were encouraging and show the potential for the clinical application of this algorithm for the detection of preterm seizures. The proposed deep learning architecture could be improved with a larger dataset of preterm EEGs with annotated seizures without the need for channel-specific seizure annotations, thus reducing the workload on clinical staff. The performance of the developed algorithm represents the current state-of-the-art in preterm seizure detection.

Chapter 7 focused on the interpretation and the proposed clinical translation of the developed deep learning models. Deep learning algorithms are widely considered to be uninterpretable black boxes, but these purely data-driven, classification optimised algorithms may offer clinical and engineering insights. This chapter disclosed three methods of probing the trained algorithms and the results are framed as comparisons between models trained for term and preterm seizure detection. The properties of trained deep learning algorithms were compared across models optimised for detecting seizures in term and preterm EEG. The compromises

and considerations associated with realising a deep learning algorithm on a mobile decision support system were also outlined in **Chapter 7**. This chapter presented work done on translating trained deep learning neonatal seizure detection algorithms to mobile platforms such as the Android tablet/phone and a custom application-specific embedded platform for use in low resource settings. The developed fully convolutional architectures offer a computationally efficient and memory efficient method of assisted interpretation while still resulting in accurate, real-time seizure classification.

8.2 Limitations and future work

The ability to utilise clinical EEG recordings of newborn infants played an important role in this research; without the necessary data this work would not have been possible. This work is however limited by the fact that all training and all testing data employed was collected from infants which had already been identified as susceptible to seizures and required further neural observation. Datasets utilised in this work did contain populations of infants which had not experienced seizure events, but all infants whose EEG was available required admission to the NICU and EEG monitoring, so these infants do not represent the general neonatal population.

This work also benefited from the EEG annotations which were provided by neonatal neurophysiology experts for the purpose of automated neonatal seizure detection algorithm development. The many experts whose time, training and knowledge went into generating seizure and non-seizure labels were vital to this research, but it must be acknowledged that this labelling is not a purely objective process. Studies have shown that highly experienced clinical experts will have some discrepancies in their seizure annotations (Stevenson et al., 2015). In an ideal scenario the annotations utilised in this work would have been the consensus of many experts' individual annotations. Due to time and labour constraints that was not always possible.

There are several avenues of exploration that arise as a result of this work. One aspect of the overall neonatal seizure detection framework that was overlooked in this thesis was the signal preprocessing and probabilistic output postprocessing routines. The preprocessing and postprocessing steps were both adopted from the previously developed SVM based ANSeR algorithm (Cork SVM algorithm). Initially, to compare the developed deep learning algorithms to the machine learning baselines it was important that the only changes made to the pipeline were in the feature extraction and classification stages. To ensure a fair evaluation of deep learning algorithm performance all other aspects of the framework were matched to the

baseline pipeline. This thesis has demonstrated the utility of deep learning algorithms and that this approach can outperform traditional feature based machine learning algorithms, so further work could explore optimising the preprocessing steps (filters, down-sampling) and the postprocessing temporal operations (moving average filter, collar, respiration adaptation).

Another parameter that was selected based on the ANSeR algorithm for comparative purposes was the EEG window size. In this work, 8-second windows of EEG were processed by the deep learning algorithm. This meant that the maximum possible size of the convolutional receptive field was limited to 8-seconds of EEG. Wider decision making windows could improve the discriminative ability of any deep learning algorithm, especially considering that other works have seen success when applying data driven classifiers to wider EEG windows (Ansari et al., 2018; Tapani et al., 2019).

The FCN architecture has been successfully leveraged throughout this work, but many other architectures could potentially be utilised to develop deep learning neonatal seizure detection algorithms. Specifically, the incorporation of temporal information into the network could further improve algorithm performance as neonatal seizure events are known to evolve over time (Patrizi et al., 2003). In the past, sequence modelling tasks were associated with recurrent architectures, but many publications have proven the utility of dilated convolutional operations and attention mechanisms for memory encoding (Bai et al., 2018; Vaswani et al., 2017).

This work focused on simple convolutional architectures which were influenced by the VGG network (Simonyan & Zisserman, 2015) - this architecture came second in the 2014 ImageNet competition (Russakovsky et al., 2015). The winning architecture in 2014 was the GoogLeNet; this architecture utilised inception blocks (Szegedy et al., 2015). These inception blocks apply multiple width convolutional filters to a layer input in parallel; this facilitates the learning of different scales of salient features. Inception networks facilitate a wider range of intermediate representations, which have proven their success in the image processing domain. Also, the introduction of the residual network architectures showed how deeper networks could be trained with the aid of residual connections (He et al., 2015); residual networks utilise shortcut connections which bypass convolutional filter operations. This direct mapping from layer input to output improves the efficiency of the learning process and allows extremely deep architectures to be trained. These more advanced architectures could prove to be beneficial in developing future neonatal seizure detection algorithms.

Moving from machine learning to deep learning has facilitated the improved utilisation of large amounts of neonatal EEG training data. Future work in this area could seek to leverage the large datasets which have been recorded and annotated as part of the ANSeR clinical trial (*ANSeR- The Algorithm for Neonatal Seizure Recognition Study*, 2014; Pavel et al., 2020; Rennie et al., 2019). The algorithms developed in this work no longer require channel-specific seizure annotations so further training can take advantage of the generalised seizure labels which are already available.

Convolutional models provide robust, reliable classifier results, but there is still a large amount of work required to interpret the internal workings of trained models, and to understand why and how these algorithms fail. Particularly when algorithms are being developed for medical decision support applications model interpretability must be emphasised. Further work in this area should focus on visualising and decoding the learned network weights through techniques such as saliency mapping and occlusion experiments (Samek et al., 2017). This research should also focus on the discovery and analysis of edge-cases; where the algorithm behaves in an unexpected way or where an unexpected input is presented to the algorithm.

8.3 References

Ansari, A. H., Cherian, P. J., Caicedo, A., Naulaers, G., De Vos, M., & Van Huffel, S. (2018). Neonatal Seizure Detection Using Deep Convolutional Neural Networks. *International Journal of Neural Systems*, *29*(04), 1850011.

ANSeR- The Algorithm for Neonatal Seizure Recognition Study. (2014). https://clinicaltrials.gov/ct2/show/NCT02160171

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *ArXiv:1803.01271 [Cs]*.

Craik, A., He, Y., & Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: A review. *Journal of Neural Engineering*, *16*(3), 031001.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *ArXiv:1512.03385 [Cs]*.

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

Pavel, A., Rennie, J., de Vries, L. S., Blennow, M., Foran, A., Shah, D. K., Pressler, R. M., Kapellou, O., Dempsey, E. M., Mathieson, S. R., Pavlidis, E., van Huffelen, A. C., Livingstone, V., Toet, M. C., Weeke, L. C., Finder, M., Mitra, S., Murray, D. M., Marnane, W. P., & Boylan, G. B. (2020). A Multicentre Investigation of a Machine Learning Algorithm for Neonatal Seizure Recognition. *The Lancet Child and Adolescent Health, published online*.

Rennie, J. M., Vries, L. S. de, Blennow, M., Foran, A., Shah, D. K., Livingstone, V., Huffelen, A. C. van, Mathieson, S. R., Pavlidis, E., Weeke, L. C., Toet, M. C., Finder, M., Pinnamaneni, R. M., Murray, D. M., Ryan, A. C., Marnane, W. P., & Boylan, G. B. (2019). Characterisation of neonatal seizures and their treatment using continuous EEG monitoring: A multicentre experience. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, *104*(5), F493–F501.

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., & Faubert, J. (2019). Deep learning-based electroencephalography analysis: A systematic review. *Journal of Neural Engineering*, *16*(5), 051001.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *ArXiv:1708.08296 [Cs, Stat]*.

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv:1409.1556 [Cs]*.

Stevenson, N. J., Clancy, R. R., Vanhatalo, S., Rosén, I., Rennie, J. M., & Boylan, G. B. (2015). Interobserver agreement for neonatal seizure detection using multichannel EEG. *Annals of Clinical and Translational Neurology*, *2*(11), 1002–1011.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going Deeper With Convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1-9.

Tapani, K. T., Vanhatalo, S., & Stevenson, N. J. (2019). Time-Varying EEG Correlations Improve Automated Neonatal Seizure Detection. *International Journal of Neural Systems*, *29*(04), 1850030.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *ArXiv:1706.03762 [Cs]*.