

This electronic thesis or dissertation has been downloaded from the King's Research Portal at <https://kclpure.kcl.ac.uk/portal/>



Individualised Clinical Neuroimaging in the Developing Brain: Abnormality Detection

MacLeod, Russell

Awarding institution:
King's College London

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without proper acknowledgement.

END USER LICENCE AGREEMENT



Unless another licence is stated on the immediately following page this work is licensed

under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International

licence. <https://creativecommons.org/licenses/by-nc-nd/4.0/>

You are free to copy, distribute and transmit the work

Under the following conditions:

- Attribution: You must attribute the work in the manner specified by the author (but not in any way that suggests that they endorse you or your use of the work).
- Non Commercial: You may not use this work for commercial purposes.
- No Derivative Works - You may not alter, transform, or build upon this work.

Any of these conditions can be waived if you receive permission from the author. Your fair dealings and other rights are in no way affected by the above.

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Thesis for the Degree of Doctor of Philosophy in Medical Imaging
Sciences and Biomedical Engineering

**Individualised Clinical Neuroimaging in the Developing Brain:
Abnormality Detection**

Russell Macleod

**EPSRC CDT in Smart Medical Imaging, Centre for the Developing Brain,
School of Biomedical Engineering and Imaging Science, Faculty of Life
Sciences and Medicine, King's College London, UK**



Abstract

Perinatal neuroanatomical structure is incredibly intricate and, at time of birth, is undergoing continuous change due to interweaving developmental processes (growth, myelination and gyrification). While there is some small variability in structure and rates of development, all follow proscribed pathways with well documented milestones. Brain injury or other disruption of these processes can result in poor neurodevelopmental outcomes or mortality, making their early identification critical to estimate, and potentially forestall, negative effects. MRI is an increasingly used method of investigating suspected neonatal encephalopathies and injuries.

Identification of these injuries and malformations is more challenging in neonates compared to adults due to the brain's continuously evolving appearance. This makes radiological review of neonatal MRI an intensive and time-consuming task which, in an ideal setting, requires a team of highly skilled clinicians and radiologists with complementary training and extensive experience. To assist this review process, some localisation method that highlights areas likely to contain tissue abnormalities would be highly desirable, as it could quickly draw attention to these locations. In addition, identifying neonates whose MRI is likely to contain some form of pathology could allow for review prioritisation.

In this thesis, I first investigated using normative models of neonatal tissue intensity for brain tissue abnormality detection. I applied voxel-wise Gaussian process (GP) regression to a training cohort of neonates with no obvious lesions, all born preterm (<37 weeks) but imaged between 28-55 weeks. Gestational age at birth (GA), postmenstrual age at scan (PMA) and sex were used as input variables and voxel intensity as the output variable. GPs output a mean value and its variance inferred from neonates within the training cohort whose demographic information most closely matched those of the prediction target. The voxel specific models were put together to form a synthesised typical image and standard deviation image derived

from the variance outputs. Z-score abnormality maps were constructed by taking the difference between neonates actual MRI and GP-calculated synthetic image and scaling by their standard deviation map. Higher Z-score map values indicate voxels more likely to contain abnormal tissue intensity. Using manually delineated masks of common brain injuries seen in a subset of neonates, these abnormality Z-score maps demonstrated good detection performance using area under the curve of receiver operating characteristic scores, with the exception of small punctate lesions.

The initial voxel-wise models had substantial false positives around the edges of the brain where there is large typical heterogeneity. I next investigated if incorporating local structural information into predictive models could improve their ability to accommodate typical anatomical heterogeneity seen across individual brains and improve the accuracy of synthetic images and abnormality detection. To achieve this, voxel intensity values in a patch surrounding the prediction target were appended to the design matrix, alongside GMA, PMA and sex. The patch-based synthetic images were able to match an individual's brain structure more closely and had lower false positives in normal appearing tissue. However, a weakness was that the centre of some larger lesions was included in the predictions (thereby classified as 'healthy' tissue), having a deleterious effect on their coverage, increasing false negatives. This was offset by much better coverage of smaller, more subtle lesions, to the extent that overall performance was higher compared to that seen in the earlier model.

I also investigated if the Z-score abnormality maps could be used to classify neonates with MRI positive brain injury from those with normal appearing brains. While many machine learning algorithms see frequent use in neuroimaging classification tasks, I opted for a logistic regression model due to its high levels of interpretability and simple implementation. Using the histograms of the Z-score abnormality maps as inputs, the model demonstrated good performance, being able to correctly identify neonates with injuries, but not those with subtle

lesions like punctate lesions, whilst minimising false identification of neonates with normal appearing brains.

To ascertain if performance could be improved, I explored multiple classification methods. Specifically, the use of other more complex classifiers (random forest, support vector machines, GP classification) and the use of a regional abnormal voxel count, that allowed localisation of lesioned tissue rather than the more global detection ability of the Z-score histograms.

Using these innovations, I investigated their application towards a specific pathology; hypoxic ischemic encephalopathy (HIE). This is a good test for the system, as HIE has high incidence rates, multiple associated lesion types and a time dependant appearance. Further, I wanted to know if, given a positive HIE diagnosis, the Z-score abnormality maps could be used to predict long-term outcomes (normal vs poor). Several models demonstrated an excellent ability to separate HIE and healthy control neonates achieving >90% accuracy, a statistically significant result even after false discovery rate (FDR) correction ($p\text{-value} < 0.05$). While the outcome prediction models achieved reasonable accuracy, >70% in multiple models, none of these were statistically significant after FDR correction.

Overall, this work demonstrates how normative modelling can be used to create individual voxel-wise / image-wise estimation of tissue abnormality for neonatal MRI across a range of gestational ages. It further demonstrates that these abnormality maps can be utilised for additional tasks, in this instance, three increasingly challenging neurological classification problems. These include the separation of neonates with and without MRI positive lesions, identification of neonates with a specific pathological condition (HIE) and prediction of long-term functional outcome (normal vs poor). Within a radiological setting, these classifications task can be considered analogous to three radiological challenges, image triage, diagnostic

detection and estimation of developmental prognosis, important for the clinical team but also infants and their families.

Declaration

The work contained within this thesis and any conclusions that have been reached are based on my own investigative research and experimentation alongside feedback from my supervisors. This project can be considered akin to a retrospective study, in that I did not conduct the data acquisition myself but applied the hypothesis and methods used within to preexisting cohorts. Where relevant, all sources that served as inspiration to sections of this work have been cited and due credit given. The primary content and conclusions of Chapter 3 and 4 have appeared as part of the PIPPI workshop of the MICCAI conference publications.

Acknowledgements

I must begin by stating the huge amount of gratitude I feel towards my supervisors, Jonathan O'Muircheartaigh, Serena Counsell and David Carmichael. Without their inspiration, wisdom, and support throughout all stages of my PhD, this thesis would not have been possible. I would also like to thank Professor David Edwards and all people associated with the centre for the developing brain (CDB) team whose insight and feedback proved invaluable throughout my research.

There are several people in particular who I would like to thank. These include my PhD thesis committee members; Emma Robinson, Flavio Dell'Acqua and Maria Deprez. Mary Rutherford and Ayse Ceren Tanritanir for their scoring of and general assistance of neonates with hypoxic ischemic encephalopathy. The PhD candidates with whom I shared my journey, including Daphna Fenchel, Irina Grigorescu, Katy Vecchiato, Ralica Dimitrova, Sofia Dall'orso and Tanya Poppe and many others. In particular, Jordina Aviles Verdera and Judit Ciarrusta, who had the, disputable, pleasure of being my desk mates and as such, due extra thanks for putting up with me. To Dafnis Batalle, Tomoki Arichi and their wider team for the welcoming atmosphere during joint lab meetings. To the Engineering and Physical Sciences Research Council (EPSRC) for funding my project. Finally, to King's College London who allowed me this fantastic opportunity.

Table of Contents

Abstract.....	- 2 -
Declaration.....	- 6 -
Acknowledgements.....	- 7 -
List of Figures	- 14 -
List of Tables	- 17 -
List of Abbreviations	- 19 -
Chapter 1. Introduction	- 22 -
1.1 Outline.....	- 22 -
1.2 Gross Brain Structure and Perinatal Development	- 25 -
1.2.1 Structure.....	- 25 -
1.2.2 Grey Matter	- 26 -
1.2.3 White Matter	- 28 -
1.3 Development	- 29 -
1.3.1 Growth	- 29 -
1.3.2 Myelination.....	- 31 -
1.3.3 Gyrification	- 32 -
1.3.4 Summary	- 35 -
1.4 Neonatal Brain Pathology	- 35 -
1.4.1 Indications of Brain Injury or Disease in Neonates	- 36 -
1.4.2 Prematurity.....	- 37 -

1.4.3	Haemorrhages, Ischemia and Infarcts	- 38 -
1.4.4	White Matter Injury	- 39 -
1.4.5	Grey Matter Injury	- 40 -
1.4.6	Cystic Lesions	- 42 -
1.4.7	Brain Structural Abnormalities and malformations	- 42 -
1.4.8	Hypoxic Ischemic Encephalopathy	- 43 -
1.4.9	Summary	- 43 -
1.5	Methods to Non-invasively Image the Brain	- 43 -
1.5.1	Clinical Neuroimaging using MRI in Neonates	- 44 -
1.5.2	T1 Weighted Imaging	- 46 -
1.5.2	T2 Weighted Imaging	- 47 -
1.5.3	Diffusion tensor Imaging	- 47 -
1.5.4	Pathological Appearance	- 49 -
1.6	The Need for Computer Assisted Medical Detection	- 50 -
1.7	Chapter Summary	- 52 -
Chapter 2. Machine Learning		- 53 -
2.1	Introduction and terminology	- 53 -
2.1.1	Model Validation	- 54 -
2.1.2	Machine Learning Approach Strategies	- 56 -
2.1.3	Measure of Performance	- 59 -
2.1.4	Scaling of Input and Output Data	- 68 -

2.2	Commonly used Machine Learning Methods	- 69 -
2.2.1	Linear Regression	- 71 -
2.2.2	Logistic Regression.....	- 71 -
2.2.3	K Nearest Neighbour (KNN).....	- 73 -
2.2.4	K-means clustering	- 75 -
2.2.5	Decision Trees and Random Forests.....	- 76 -
2.2.6	Markov Random Fields (MRF)	- 77 -
2.2.7	Support Vector Machines (SVM)	- 79 -
2.2.8	Neural Networks	- 80 -
2.2.9	Gaussian Processes	- 83 -
2.2.10	Model Parameter and Hyper-Parameter Optimisation.....	- 85 -
2.3	Existing use of Machine Learning within Neuroimaging	- 86 -
2.3.1	Lesion Detection: Segmentation of Diseased Tissue.....	- 86 -
2.3.2	Categorical Disease Classification.....	- 92 -
2.3.3	Machine Learning Methods Review	- 98 -
2.3.4	An alternative approach – Normative modelling.....	- 100 -
2.3.5	Conclusions.....	- 101 -
2.3.6	Decisions.....	- 103 -
2.4	Project goals	- 104 -
2.6	Hypothesis and Aims	- 105 -
	Chapter 3. Automatic Detection of Neonatal Brain Injury on MRI	- 108 -

3.1	Introduction	- 108 -
3.2	Methods.....	- 113 -
3.2.1	Sample and dataset.....	- 113 -
3.2.2	Pre-processing.....	- 116 -
3.2.3	Modelling Voxel Intensity Values using Gaussian Processes	- 117 -
3.2.4	Detecting Pathological Tissue as Outliers of Typical Development	- 121 -
3.3	Results	- 122 -
3.4	Discussion	- 127 -
3.5	Conclusion.....	- 131 -
Chapter 4. Detection of Injury and Automated Triage of Preterm Neonatal MRI using Patch-Based Gaussian Processes..... - 133 -		
4.1	Introduction	- 134 -
4.2	Methods.....	- 136 -
4.2.1	Sample and dataset.....	- 136 -
4.2.2	Pre-processing.....	- 139 -
4.2.3	Base Gaussian Process Tissue Intensity Model	- 139 -
4.2.4	Patch Extension to the Gaussian Process Tissue Intensity Model	- 140 -
4.2.5	Detecting Pathology	- 142 -
4.2.6	Automated Clinical Image Triage.....	- 143 -
4.3	Results	- 144 -
4.3.1	Base Gaussian Process Tissue Intensity Model	- 144 -
4.3.2	Patch Based Gaussian Process Tissue Intensity Model	- 145 -

4.3.3	Automated Triage of Abnormal Images	- 149 -
4.4	Discussion	- 150 -
4.5	Conclusion.....	- 159 -
Chapter 5. Automatic Detection of Hypoxic Ischemic Encephalopathy using Outlier Detection		- 160 -
5.1	Introduction	- 161 -
5.2	Methods.....	- 167 -
5.2.1	Sample and dataset.....	- 167 -
5.2.2	Pre-processing.....	- 169 -
5.2.3	Generating Subject Specific Predicted and Outlier Z-score Maps	- 170 -
5.2.4	Injury Localization.....	- 172 -
5.2.6	Classification Algorithms	- 174 -
5.2.7	Estimating Model Performance	- 178 -
5.3	Results	- 180 -
5.3.1	Synthesized Typical Brain Images and Z-scores	- 180 -
5.3.2	Classification of Hypoxic Ischemic Encephalopathy	- 182 -
5.3.3	Classification of Positive and Negative Outcomes.....	- 185 -
5.4	Discussion	- 188 -
5.5	Conclusion.....	- 198 -
Chapter 6. Thesis Discussion and Conclusions		- 199 -
6.1	Thesis Overview.....	- 199 -
6.2	Limitations and Improvement	- 206 -

6.3	Final Remarks.....	- 208 -
Chapter 7. Bibliography and Citations..... - 210 -		
Chapter 8. Appendices - 246 -		
8.1	Chapter 3 Publication - Automatic Detection of Neonatal Brain Injury on MRI-	246
	-	
8.2	Chapter 4 Publication- Detection of Injury and Automated Triage of Preterm Neonatal MRI Using Patch-Based Gaussian Processes.....	- 256 -
8.3	Chapter 5 Raw Data Tables.....	- 267 -

List of Figures

Figure 1.1 Cortical grey matter areas of the brain

Figure 1.2 Deep grey matter areas of the brain

Figure 1.3 White matter areas of the brain

Figure 1.4 Brain volumes at different gestational ages

Figure 1.5 Total brain volume charts over the perinatal period

Figure 1.6 Single cell myelination

Figure 1.7 Patterns of myelination

Figure 1.8 Gyrification over the perinatal period

Figure 1.9 Sulcus diagonalis structural combinations

Figure 1.10 T1w images at different ages

Figure 1.11 T2w images at different ages

Figure 1.12 Images of neonatal pathology

Figure 2.1 Binary confusion matrix

Figure 2.2 Multi-class confusion matrix

Figure 2.3 Receiver operating characteristic and precision recall curves

Figure 2.4 Sigmoid Function

Figure 2.5 K nearest neighbour

Figure 2.6 K means clustering

Figure 2.7 Decision tree classifier

Figure 2.8 Markov random field (undirected graph)

Figure 2.9 Support vector machines

Figure 2.10 Basic neural network architecture

Figure 2.11 Simplified Gaussian process

Figure 3.1 Preprocessing and GP performance

Figure 3.2 Abnormality detection performance

Figure 3.3 Individual pathology detection performance

Figure 3.4 Missed pathologies and smoothing of the predicted images

Figure 4.1 Example 5x5x5 patch model

Figure 4.2 T1w/T2w predicted images

Figure 4.3 Improved patch-based model performance over the base model

Figure 4.4 Improved ROC performance in patch-based models

Figure 5.1 Images of the 85-region atlas used to create the regional abnormal voxel count

Figure 5.2 Flow chart of the different classifiers used in this chapter

Figure 5.3 Patch-based prediction of HIE neonate with enlarged ventricles

Figure 5.4 HIE vs HC using histograms

Figure 5.5 HIE vs HC histogram permutation testing

Figure 5.6 HIE vs HC using regional abnormal voxel count

Figure 5.7 HIE vs HC regional abnormal voxel count permutation testing

Figure 5.8 HIE normal vs poor outcomes using histograms

Figure 5.9 HIE normal vs poor outcomes using histograms permutation testing

Figure 5.10 HIE normal vs poor outcomes using regional abnormal voxel count

Figure 5.11 HIE normal vs poor outcomes using histograms permutation testing

Figure 5.12 Typical HIE lesion detection

Figure 5.13 Pathology detection at different thresholds

Figure 5.14 Typical HIE lesion detection with emphasis on ADC

Figure 5.15 Neonates with difficult to predict outcomes

List of Tables

Table 2.1 Machine learning algorithms

Table 3.1 Acquisition parameters

Table 3.2 Cohort information

Table 3.3 The mean specificity, sensitivity and precision of models and a threshold of $Z > 3$

Table 3.4 Z-Score threshold comparison

Table 3.5 Area under curve of receiver operating characteristic for each of the preprocess methods and raw Z-score against the individual raters

Table 4.1 Acquisition parameters

Table 4.2 Cohort Information

Table 4.3 AUC for all models on the 1st and 3rd iteration

Table 4.4 Sensitivity and specificity for the 36 images with manual segmentations from the initial preterm dataset

Table 4.5 Sensitivity and specificity for the 20 images with manual segmentations from the second brain injury dataset

Table 4.6 Neonates from the initial preterm cohort that were correctly identified by the different 7x7x7 surface patch models

Table 4.7 Neonates from the second brain injury cohort that were correctly identified by the different 7x7x7 surface patch models

Table 5.1 Acquisition parameters

Table 5.2 Cohort information

Table 5.3 HIE cohort with outcome information

Table 5.4 The regions present in the dHCP based atlas and their different subdivisions

Table 5.5 Classification methods used in this chapter and their parameters

Table 5.6 Regional abnormal voxel count optimal threshold testing

Table 8.1 Sensitivity and specificity scores for each HC vs HIE classification models using raw Z-score histogram from each modality

Table 8.2 Permutation testing of the HC vs HIE classification models using raw Z-score histograms from each modality

Table 8.3 Sensitivity and specificity scores for each HC vs HIE classification models using optimised regional abnormal voxel count from each modality

Table 8.4 Permutation testing of the HC vs HIE classification models using optimised regional abnormal voxel count from each modality as input

Table 8.5 Sensitivity and specificity scores for each HIE normal vs poor outcome classification models using raw Z-score histogram from each modality

Table 8.6 Permutation testing of the HIE normal vs poor outcome classification models using raw Z-score histograms from each modality

Table 8.7 Sensitivity and specificity scores for each HIE normal vs poor outcome classification models using raw Z-score histogram from each modality

Table 8.8 Permutation testing of HIE normal vs poor outcome classification models using optimised regional abnormal voxel count from each modality

List of Abbreviations

AD Alzheimer's disease

ADC Apparent diffusion coefficient

AI Artificial intelligence

ARD Automatic relevance determination

AUC Area under the curve

BGT Basal ganglia and thalamus

CAD Computer aided design

CH Cerebellar haemorrhage

CPVL Cystic periventricular leukomalacia

CSF Cerebral spinal fluid

CT Computed tomography

DEHSI Diffuse excessive high signal intensity

DTI Diffusion tensor imaging

EOAD Early onset Alzheimer's disease

FDR False discovery rate

FLAIR Fluid-attenuated inversion recovery

GM Grey matter

GA Gestational age at birth

GLH Germinal matrix haemorrhage

GP Gaussian process

GPC Gaussian process classifier

HC Healthy control

HPI Haemorrhagic parenchymal infarct

ICA Independent component analysis

IVH Inter-ventricular haemorrhage

KNN K nearest neighbour

LDA Linear discrimination analysis

LIN-SVM Linear kernel support vector machine

LOAD Late onset Alzheimer's disease

LR Logistic regression

MAE Mean absolute error

MCI Mild cognitive impairment

ML Machine learning

MPRAGE Magnetization-prepared rapid gradient-echo

MRF Markov random field

MRI Magnetic resonance imaging

MS Multiple sclerosis

MSE Mean square error

PCA Principal component analysis

PET Positron emission tomography

PMA Postmenstrual age at scan

PRC Precision recall curve

PVL Periventricular leukomalacia

RAVC Regional abnormal voxel count

RBF-SVM Radial basis function kernel support vector machine

RF Random forest

RMSE Root mean square error

ROC Receiver operating characteristic

SVM Support vector machine

T1w T1 weighted

T2w T2 weighted

TE Echo time

TR Repetition time

WM White matter

Chapter 1. Introduction

1.1 Outline

The perinatal period (25 to 44 weeks gestational age) is a time of rapid neuroanatomical development involving multiple intersecting processes (Philip, 2005). For the postnatal brain, the period encompasses both neonates born preterm, that is neonates born before 37 weeks gestational age (GA), and term birth, those born between 37- and 42-weeks GA. Premature or late birth (after term birth age range), injury, pathology, or other forms of neurodevelopmental insult can result in long lasting or permanent effects on an individual's later outcome as a child and adult (Anand & Scalzo, 2000). These injuries and insults can be caused by, for example, genetic conditions, maternal disease, prenatal injury, complications related to the pregnancy and birth, and post-natal injury. Identification and monitoring of these perinatal brain insults can potentially inform interventions or other treatment crucial to assuage negative effects.

Where necessary, clinical neuroimaging is a widely used adjunct for both maternal and neonatal monitoring and to assist the diagnostic process (Baysinger, 2010). It allows in vivo examination of a neonate, and more specifically here their brain, either in-utero or postnatally. Two particular imaging techniques that see high use in the neonatal setting are ultrasound and magnetic resonance imaging (MRI) (M. Rutherford, Biarge, et al., 2010; Whitworth et al., 2015). While ultrasound sees ubiquitous use for routine in-utero assessments, MRI has seen increasing use for diagnosis, identification and extent estimation of pathological brain conditions both pre- and post-natally. MRI has high resolution, high signal-to-noise and allows for different forms of acquisitions giving multiple tissue contrasts. Its use in perinatal brain examinations is increasingly mature with well-established radiological guidelines (M. A. Rutherford, 2002).

Even with the support provided by clinical imaging, identifying specific neurological conditions or injuries is a challenging task. Perinatal brain pathologies are heterogeneous and individual brain conditions can have multiple presentations. In addition to this, and more visually obvious, is the rapid change in brain structure and tissue occurring during the brain's development. Not only does this cause significant but predictable alterations in the appearance of the neonatal brain over time but the rate at which these changes occur can vary between individuals. This reflects the complex nature of the brain itself. In adults, the brain has a highly complex structure with heterogeneity in the exact tissue border location and cortical folding pattern, although the gross position, shape and volume is generally consistent and predictable (Bethlehem et al., 2022).

In clinical interpretation by a radiologist, there arises the potential of neuroanatomical features being misidentified without extensive specialist training. Specifically, tissues, structures or developmental patterns may be incorrectly identified as abnormal when they are in fact healthy and typical. For the same reason, this could lead to some real brain abnormalities being obfuscated and pathological neonates incorrectly identified as being healthy. The large variety of neurological injuries, deformities and pathologies that may be present over the perinatal period in turn have a large variety in appearance, volume and location. The result is that accurate knowledge of typical developmental processes, their representation on MRI and the timeframe of their occurrence is crucial to correctly detect and identify any form of pathological presence.

For this reason, all clinical staff involved in the radiological review process undergo extensive training. During examination of a neonate's MRI and their associated clinical and demographic information, at least one member of staff should be a senior radiologist with extensive experience, although this may not always be possible (NHS Clinical Guideline 2017). In an ideal situation, it should be undertaken by a multi-disciplinary team of clinicians with

complementary training and experience. A review conducted in this manner allows for a diagnostic consensus of any neurological abnormality and its clinical relevance to be reached. Even then, there can still be reporting inconsistencies of neurological conditions between different individuals or clinical groups, especially in estimations of their severity and extent (Bogner, 2018).

Outside of specialised hospitals this clinical expertise may not be readily available. In this context, objective measures that isolate or identify whole images, or areas within those images, containing abnormal intensities or features could assist in this extremely challenging area. Computer aided detection, in particular, could reduce the burden on clinical staff by assisting with identification and improving consistency when estimating pathological extent, while not replacing clinical expertise (Doi, 2007; X. Liu et al., 2019). This is demonstrated by the rapid development and expand usage of computer aided diagnosis in adults (Yanase & Triantaphyllou, 2019).

In this thesis, I will address the challenges of detecting neonatal pathologies and abnormalities. Machine learning techniques will be combined with expert clinical knowledge of changing neonatal brain's structure and pathology, to initially develop a generalised abnormality detection system for the neonatal brain. Next, I will address the use of machine learning in diagnosis of a clinically variable neonatal brain disorder, hypoxic ischemic encephalopathy. Finally, I will address one of the most challenging problems in clinical neonatal neurology, prognosis in individual neonates. My primary aim would be to assist staff in identification of neonates with brain injury, mapping the extent of these abnormalities and exploring the use of these techniques to estimate likely long-term outcome of brain injury in individual neonates. There is within the medical research community an opinion that can be summarised by a quote by Dr Bertalan Mesko: "Artificial intelligence (AI) will not replace medical professionals. However, those professionals that use AI will replace those that don't".

The remaining contents of this chapter will be as follows. An overview of neonatal brain tissue and structure along with the changes they undergo thanks to the various developmental processes during this period. Information of the causes and descriptions of the different pathologies that are commonly seen in neonates. An introduction to the different imaging techniques available in hospitals and that can be applied to neonates. A description of both normal and pathological tissue from that neuroimaging context. Finally, a brief reiteration of the goals of this thesis based on the information covered below.

1.2 Gross Brain Structure and Perinatal Development

Before any detection of abnormality can be attempted, it is necessary to understand both the brain structure and the wide range of developmental processes occurring during the perinatal period: growth, myelination and gyrification (Bozek et al., 2018; Schuh et al., 2018). Understanding these developmental processes, their appearance, trajectories and the interactions between them is crucial for characterising “typical” development and therefore quantifying atypical development.

1.2.1 Structure

At the highest level, the lobes (frontal, parietal, occipital, temporal), as well as the cerebellum, brain stem and pons are all present at birth (Holland et al., 2014). The ventricular system, composed of the lateral, 3rd, 4th ventricles, cerebral aqueduct and choroid plexus, is complete albeit with a higher volume relative to the rest of the brain compared to adults (Brouwer et al., 2012). Also present are the brain fissures, separations between the lobes and hemispheres, along with many of the gyri and sulci, the ridges and grooves of the cortical surface Fig 1.1 (Chi et al., 1977; Hill et al., 2010). Many of these folds, though not all, will be present at birth, albeit shallower than those seen in adults as gyrification is ongoing from the 2nd to 3rd trimester.

1.2.2 Grey Matter

The exterior surface of the brain is covered by several thin protective layers including the dura, arachnoid and pial (Kolb & Whishaw, 2021). Lying just beneath these, and constituting the first layer of the brain proper, is a thin section of tissue known as the cortical grey matter that is responsible for much of the brain's functional operation, Fig 1.1 row B.

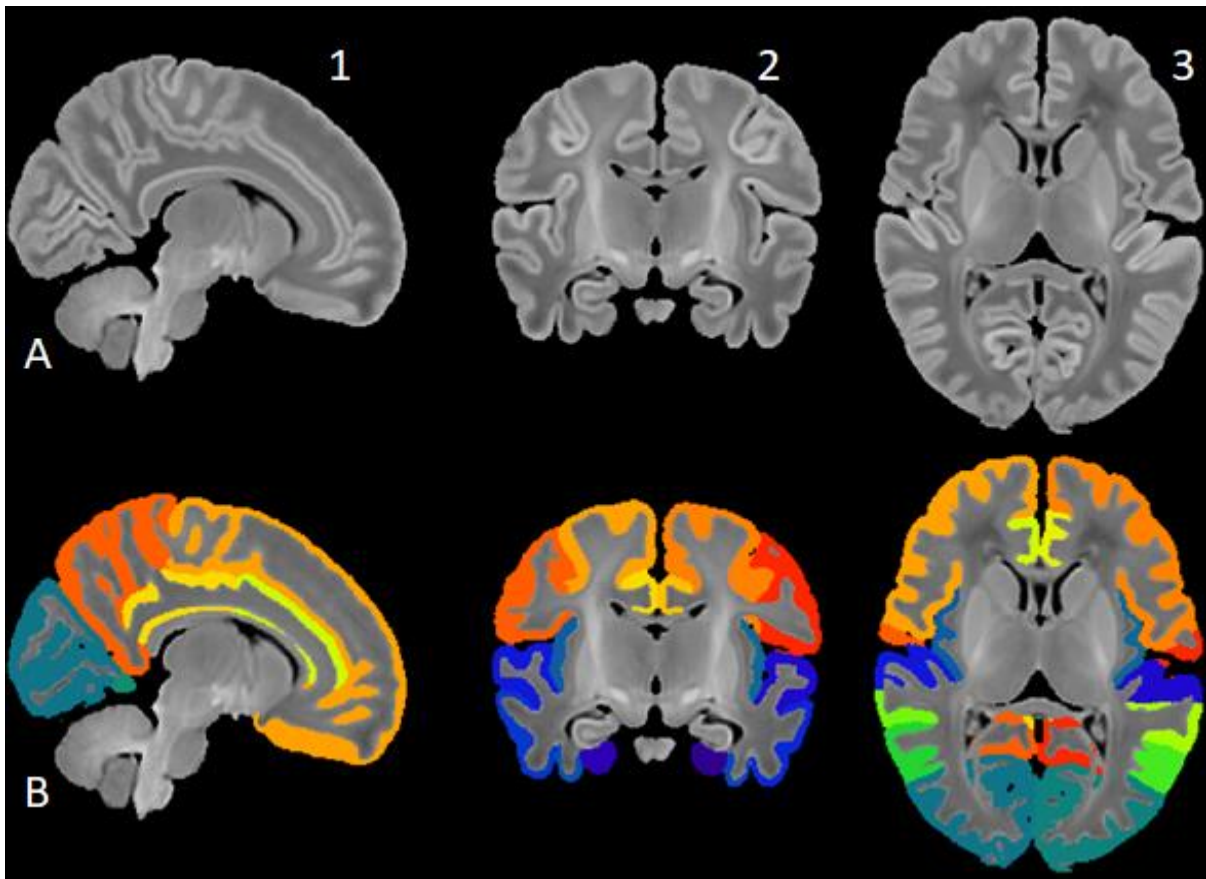


Figure 1.1 Cortical grey matter areas of the brain. Row A shows a T1w week 40 post menstrual age brain template image and B shows different coloured cortical GM regional overlays in the sagittal, coronal and axial planes, columns 1, 2 and 3 respectively.

Some of these functions, such as monitoring of sensory input or control of motor functions, are known to have a strong association with local areas of cortex. Other functions, such as memory, skill learning and emotional control, often rely on a complex interaction between spatially disparate but functionally connected areas (Hampson et al., 2010; Kim et al., 2011). To further complicate matters, areas of cortex and subcortex (e.g. the thalamus) can be responsible for relaying, combining and controlling information, both between different areas of the brain and

the rest of the central and peripheral nervous system (Herrero et al., 2002). This means that interruption or injury to one area can have down- or upstream consequences on whole brain function.

The other major structures composed primarily of grey matter include the basal ganglia (caudate nucleus, putamen, globus pallidus), thalamus and hypothalamus. Much like cortical grey matter, these structures are highly involved in the functional operation and control of the brain. Specifically, these areas act within circuits for interchange of information responsible for motor learning and control, emotion, behaviour and executive decision-making Fig 1.2 (Lanciego et al., 2012).

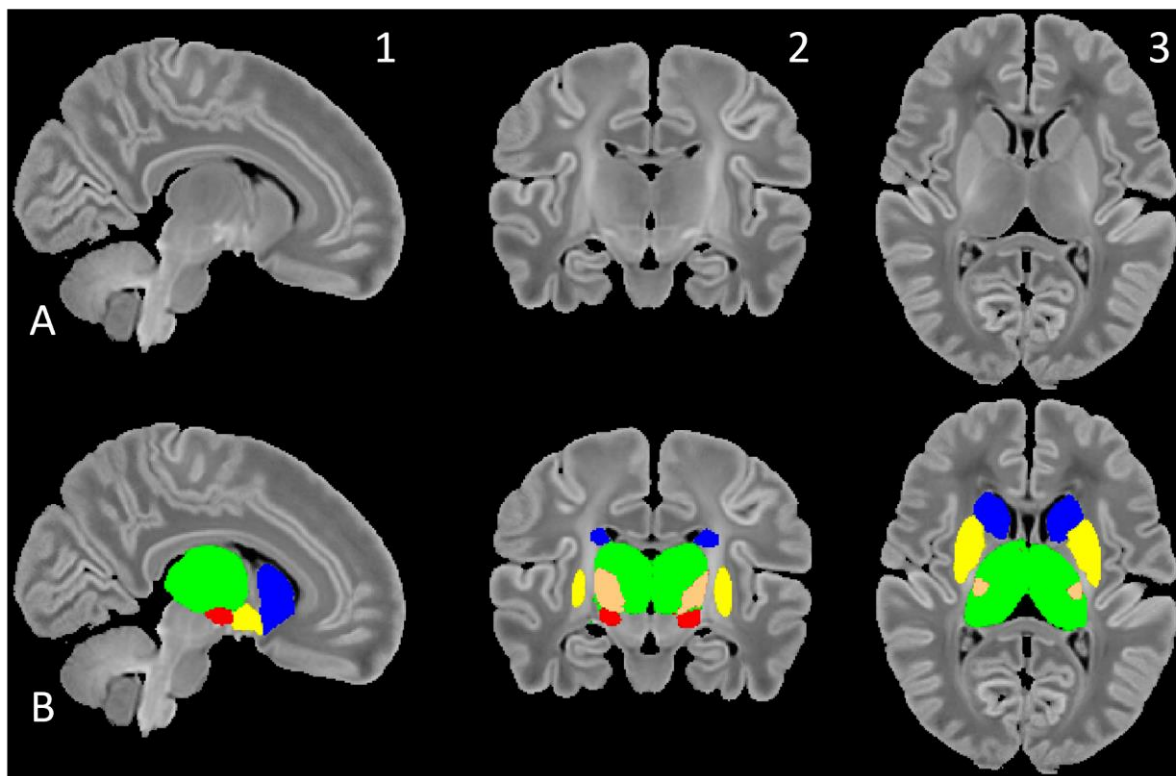


Figure 1.2 Deep grey matter areas of the brain. Row A shows a T1w week 40 post menstrual age brain template image and B shows different coloured deep GM regional overlays in the sagittal, coronal and axial planes, columns 1, 2 and 3 respectively. The structures shown are the caudate nucleus (blue), lentiform nucleus (yellow), subthalamic nucleus (red) as well as the thalamus (green) and the ventrolateral nuclei of the thalamus (bronze).

1.2.3 White Matter

The white matter is composed primarily of neuronal axons supported by oligodendrocytes, which myelinate these axons allowing increased rates of inter-neuronal information transmission throughout spatially separated areas of the brain and between cortical and subcortical areas (Kolb & Whishaw, 2021). Several well-known major white matter tracts exist, including the corpus callosum, internal capsule limbs, external capsule, cingulum, cortico-spinal tract, fronto-occipital fasciculus, inferior longitudinal fasciculus, superior longitudinal fasciculus, spino-thalamic tract, uncinate fasciculus and corona radiata. An overview of white matter areas of the brain is shown in Fig 1.3.

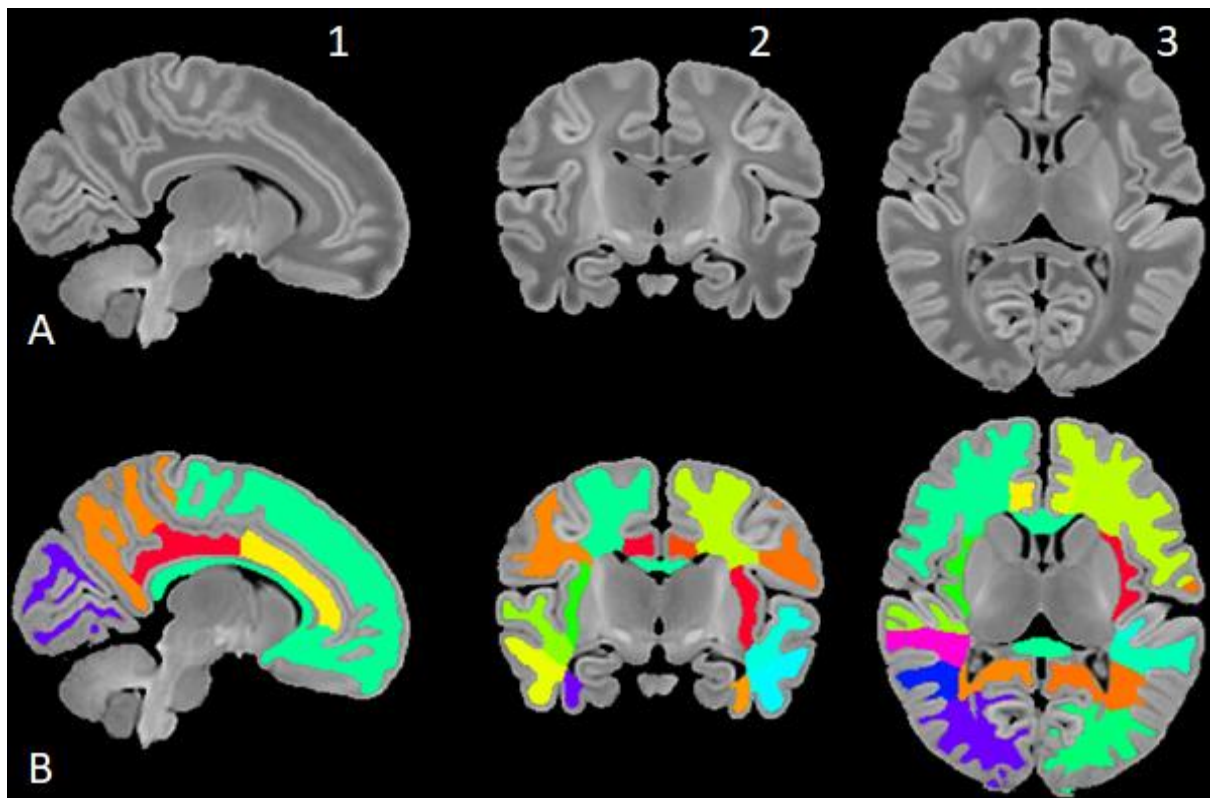


Figure 1.3 White matter areas of the brain. Row A shows a T1w week 40 post menstrual age brain template image and B shows different coloured WM regional overlays in the sagittal, coronal and axial planes, columns 1, 2 and 3 respectively.

1.3 Development

Unlike in adults, where the structure of the brain changes slowly over the scale of years, the appearance of the neonatal brain changes rapidly with age, depending on several distinct and interwoven developmental processes. These processes (growth, myelination and gyrification) are already underway by term birth, having begun in-utero during the foetal stages.

1.3.1 Growth

The most obvious developmental change seen in the neonatal brain is its rapid growth. Beginning at gestation, brain volume increases substantially, continuing up to birth and then postnatally to the early teens, where total brain volume reaching its maximal value at age 12 (Bethlehem et al., 2022). Although initially extremely rapid, this rate of growth decreases over time as an individual matures (Groeschel et al., 2010; Knickmeyer et al., 2008). A visualisation of brain sizes at different ages is shown in Fig 1.4 (Batalle et al., 2018).

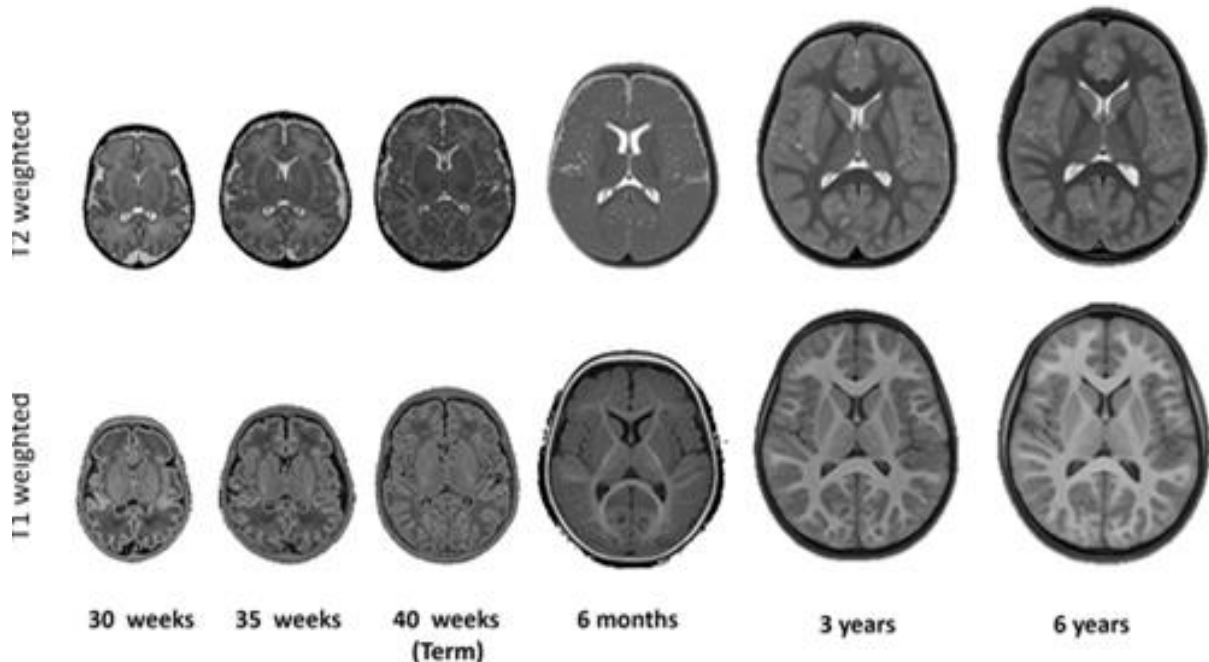


Figure 1.4 Brain volumes at different gestational ages. From left to right, cortical surface reconstructions at 30 weeks, 35 week, 40 weeks post menstrual age, 6 months, 3 year and 6 years. Adapted from Batalle et al., 2018.

While there are many contributors to increasing postnatal brain volume, it is predominantly driven by two other developmental processes: the proliferation of non-neuronal glia, leading to myelination of the white matter, as well as with dendritic and synaptic proliferation. The myelination process is covered in section 1.3.2, but, in brief, refers to ensheathing of neuronal axons with myelin (Barkovich, 2000; Welker & Patton, 2012). This increases axonal diameter inflating white matter volume. Dendritic arborisation and synaptic proliferation describe first the extrusion of dendritic extensions from the cell body; followed by their interconnection with other dendrites and the axon terminals from neighbouring neurones (Vasung et al., 2019).

A more quantitative view of this process can be given by measuring of the total brain volume (TBV) at different stages of development. Typical in-utero TBV increases from approximately 100cm^3 at 26 weeks GA to up to 300cm^3 at 36 weeks GA by Clouchoux et al., 2012. This was replicated, and refined, by later work from Jarvis et al., 2019 who found TBV of 27.5cm^3 at 19 weeks GA, 101.5cm^3 at 26 weeks GA and 219cm^3 at 36 weeks GA. Postnatally, 37-44 weeks GA, TBV demonstrated an increase in volume to 425.39cm^3 (44-weeks) which rose to

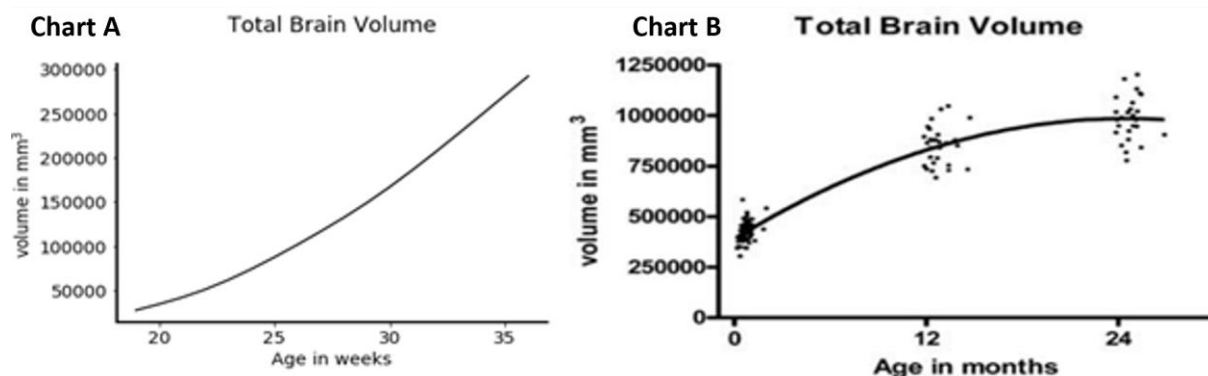


Figure 1.5 Total brain volume charts over the perinatal period. Chart A, created from data in Jarvis et al., 2019, shows the perinatal TBV from 18 weeks to 40 weeks gestational age. Chart B, taken directly from and Knickmeyer et al., 2008, shows the TBV from birth until 26 months. As an interesting side note, it is quite satisfying to see how the adult TBV appears as if it would naturally continue from the neonatal chart if both were applied to the same axis.

855.54cm³ after 1 year and to 983.87cm³ after two years in term neonates (Groeschel et al., 2010; Knickmeyer et al., 2008). Therefore, the prenatal and postnatal growth rates far exceed those seen after 12 months. Two brain growth charts based on information taken from Jarvis et al., 2019 and taken directly from Knickmeyer et al., 2008 can be seen in Fig 1.5

1.3.2 Myelination

Myelination is the process involving the proliferation, growth and migration of highly specialised microglial cells called oligodendrocytes (McArdle et al., 1987). Oligodendrocyte formation is unique, they develop so that their cell bodies wrap around and encase the neuronal axons, shown in Fig 1.6 (Xin & Chan, 2020). This is done in such a way as to allow small, periodic gaps of exposed axon in the myelin sheath, nodes of Ranvier (Kolb & Whishaw, 2021). These nodes increase the speed of information transmission along the axon (saltatory conduction), allowing for faster communication between cells and, by extension, the different regions of the brain (Barkovich, 2000).

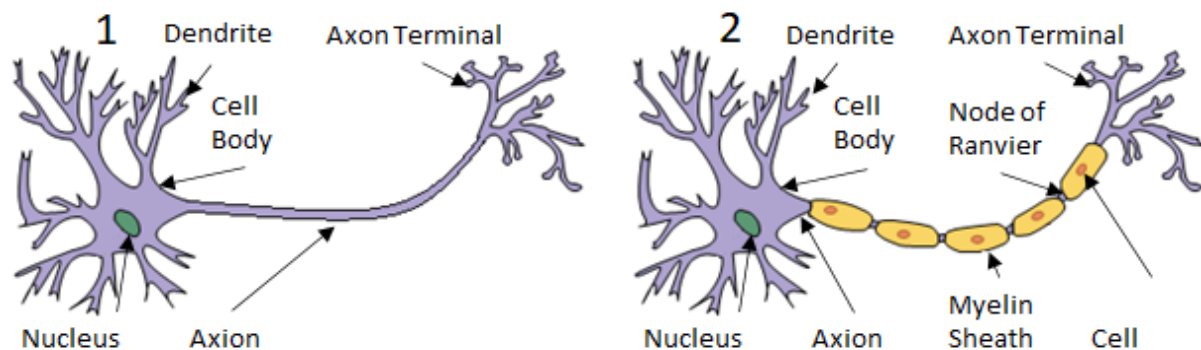


Figure 1.5 Single cell myelination. This figure shows a single neurone before (column 1) and after (column 2) myelination.

Small amounts of myelinated oligodendrocytes can be detected throughout the brain as early as 14 weeks GA, although these are isolated and not widespread (de Graaf-Peters & Hadders-Algra, 2006). The first truly myelinated area of the brain is present at around 20 weeks GA in the globus pallidus and pallid thalamic fibres. From here, several patterns of myelination have

been documented, although as with other developmental processes their rate of progress and exact path may differ between individuals (Branson, 2013; Welker & Patton, 2012).

As a general rule, myelination of cortical and subcortical structures progresses along two major axes, inside out and posterior to anterior, following on from myelination of the neurones in the spinal tracts, Fig 1.7. It also moves from the brain's centre to its periphery and, in the cortical areas, from the sensory to motor regions (Branson, 2013). Having begun in-utero, myelination reaches a mature, adult like appearance at around 18 months. However, myelin's plasticity means that development, refinement and repair continue throughout life (de Graaf-Peters & Hadders-Algra, 2006; Xin & Chan, 2020).

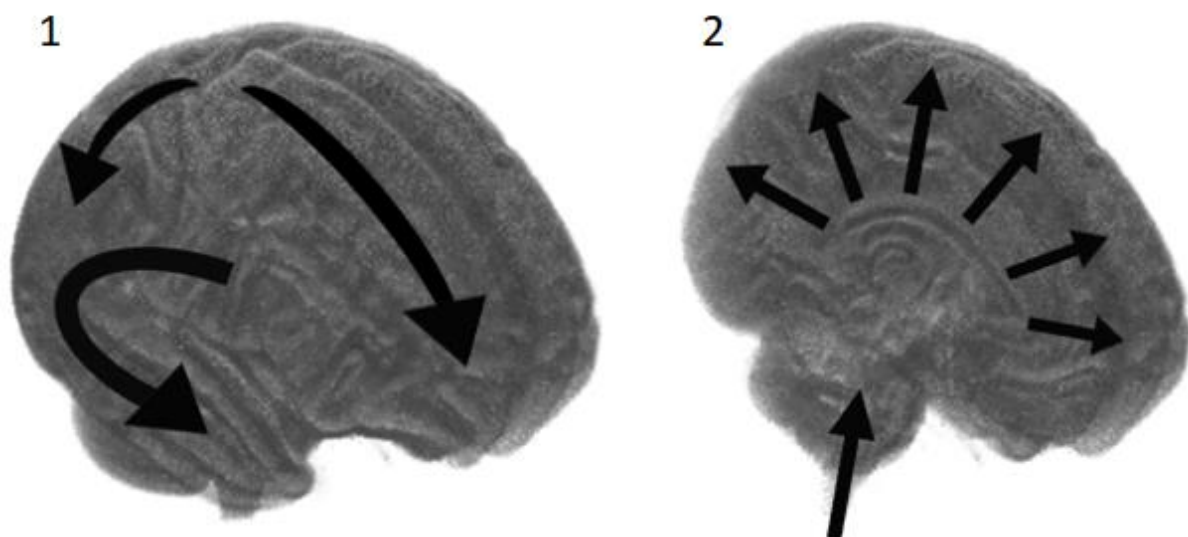


Figure 1.6 Patterns of myelination. Column 1 shows the external pattern of myelination around the cortical surface, column 2 shows the internal myelination pattern following on from myelination of the spinal tract.

1.3.3 Gyrification

Another developmental change is the gyrification process. Over the perinatal period, the smooth, largely homogeneous cortical surface undergoes an intricate folding process to form a complex and heterogeneous pattern of gyri and sulci seen in later life, see Fig 1.8 (G. Li et al., 2014; Rajagopalan et al., 2011). Fissures and sulci develop at different times beginning at 12

weeks with the interhemispheric fissure. The remaining primary sulci form continuously throughout in-utero development and the secondary just before, during or just after birth. An individualised final cortical folding pattern is in around the first year of life after gestation (G. Li et al., 2014; Nishikuni & Ribas, 2013). The later developed gyri are markedly heterogeneous between individuals.

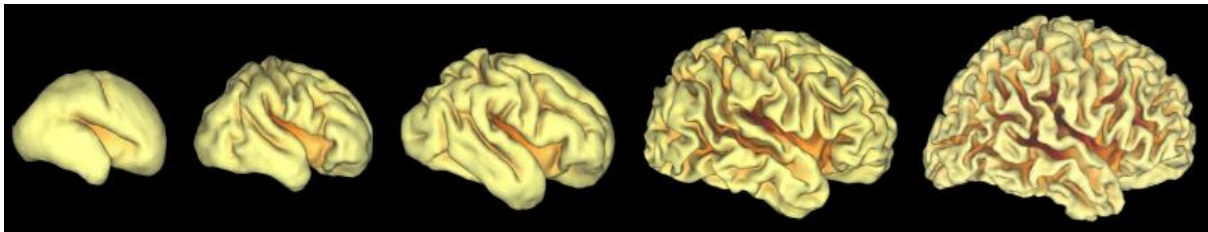


Figure 1.7 Gyrification over the perinatal period. From left to right, example cortical folding patterns seen at 25, 30, 33, 39 weeks gestational age and an adult. Derived from images provided by the Van Essen Lab (Washington University in St. Louis).

At the earlier stages of development, this outer surface appears smooth with little to no gyri or sulci. This smooth surface limits the number of neurons and supporting cells that can be present on the cortical surface. Gyrification has the effect of increasing the overall surface area of the brain, allowing for increased neuronal concentration (Chi et al., 1977).

The mechanisms behind gyrification are still a highly debated subject. Traditionally four theories have been posited as the primary contributors to cortical folding: skull constraint, axonal tension, cortical growth and programmed overgrowth of certain areas (Garcia et al., 2018). Of these, they identified heterogeneous cortical growth was linked to formation of the primary gyri, with additional contribution via different tangential growth, a view supported by work from (Garcia et al., 2018; Ronan & Fletcher, 2015). In these instances, folding is caused by internal pressure and constrained by an elastic effect of the tissue, its relative stiffness and cortical thickness.

The earliest identifiable sulci and fissures are the interhemispheric fissure and callosal sulcus at 12 weeks, the hippocampal sulcus at 15 weeks, the lateral sulcus/sylvian fissure, circular insular sulcus, olfactory sulcus, calcarine sulcus and parietoccipital sulcus at 17 weeks (Nishikuni & Ribas, 2013). This continues with the gyrification of the cingulate sulcus, central sulcus and orbital sulcus at around 19-21 weeks, all of which will be visible in “normal appearing” in-utero scans (Rajagopalan et al., 2011; Li et al., 2014). This is followed by formation of the majority of the major gyri and sulci between 24-30 weeks GA. The exception to this being the inferior and transverse temporal sulcus, which form later at 31 and 33 weeks respectively (Li et al., 2014; Nishikuni & Ribas, 2013). While many of an individual’s major gyri and sulci will be in place at this time, they have yet to reach their complete sulcal depth, and gyral peaks. Formation of the secondary gyri and sulci begins at 38 weeks, roughly in line with term birth. Complete gyrification is achieved by 52 weeks GA, having reached its final intricate and heterogeneous folding pattern (G. Li et al., 2014; Rajagopalan et al., 2011).

Many of the primary gyri and sulci cortical folding patterns have a general form, with the majority of neonates having similar sulci and gyri formations across individuals. However, the size, orientation and exact position of these patterns is highly heterogeneous (Meng et al., 2018; Zilles et al., 1997). Additionally, some gyri and sulci may or may not be present, extend for different lengths and intersect one another at different points across different individuals. An example of this can be seen in the sulcus diagonalis, a small formation located anterior to the ascending ramus of the lateral fissure. In 48.75% of subjects, the sulcus diagonalis may not be present or appear so small as to be indiscernible, sometimes resulting in its misidentification, demonstrated in Fig 1.9 (Sprung-Much & Petrides, 2018).

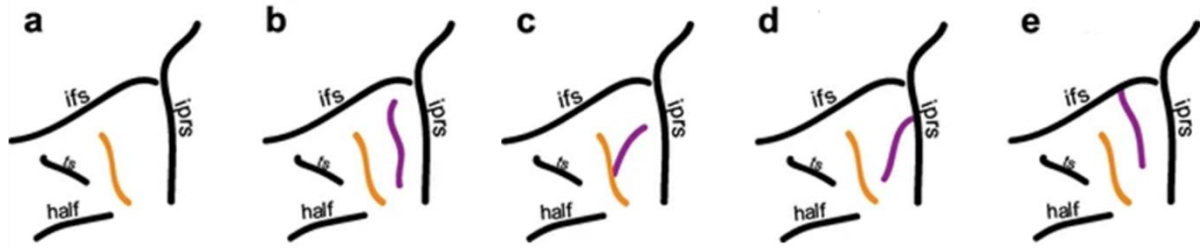


Figure 1.9 Sulcus diagonalis structural combinations. Columns a to e demonstrate several possible combinations of sulcus diagonalis (purple) and the ascending ramus of the lateral fissure (orange). The positions of the inferior frontal sulcus (ifs), inferior precentral sulcus (iprs), triangular sulcus (ts) and horizontal ramus of the lateral fissure (half) are also shown for context. Taken from Sprung-Much & Petrides, 2018.

1.3.4 Summary

An important consideration for all of these developmental processes, is that their rate of progress changes over time, and that there can be marked heterogeneity of anatomy between individuals of the same age. The level of these deviations also increases with age and type, for example the tertiary sulci, appearing ≈ 38 -weeks GA, being more variable than earlier primary and secondary sulci. While some deviation from these prescribed developmental trajectories can be considered typical, anatomical development can also be influenced by environmental factors, congenital conditions, and morbidity (Boardman & Counsell, 2020; Dimitrova, Pietsch, et al., 2021; Gilmore et al., 2010).

1.4 Neonatal Brain Pathology

One of the most significant challenges to be addressed in this thesis is accounting for the large variety of potential neurological abnormalities that can be present in the neonatal brain, specifically identifying what is and is not typical structure. This is further complicated by the variation in position, size and shape that may be present within even a single type of pathology (M. A. Rutherford, 2002). Additionally, while many pathologies have an observable effect on the brain, others may only be categorizable via qualitative assessment of clinical information.

That is, they have some diagnosable effect, such as a change in wakefulness, feeding difficulties or blood oxygenation level, but little to no detectable effect on brain physiology. Even with those that have an observable effect, it is still possible to miss small injuries or tissue abnormalities with relatively minor effects on brain tissue appearance. This combination of factors highlights why neonatal clinical and radiological experience is so critical to the identification process.

1.4.1 Indications of Brain Injury or Disease in Neonates

Most observable abnormalities can be subdivided into either primary tissue injury, tissue injury secondary (in time) to an injury or disease, or structural malformation. Common injuries seen in neonates include haemorrhages, cysts, focal and diffuse tissue lesions and bacterial, viral and fungal infection (M. A. Rutherford, 2002). These can be the result of an external effect such as prior injury, premature birth, difficult delivery, post-natal trauma or infection. Genetic and congenital factors may be further affected or enhanced any injuries that may be present (Boardman et al., 2014; Kelly et al., 2019). Structural aberrations include enlarged ventricles, increased extra cortical space, under- or over-sized brain regions, cortical malformations (such as polymicrogyria) and vascular malformations. They can be present as a secondary response to an initial injury or may be a distinct medical condition in and of themselves.

With the exception of prematurity and known pre-existing conditions, the presence of pathology or structural abnormality can only be indicated by assessment of these unusual clinical characteristics by medical professionals. These include clinical scores recorded directly from the neonate at birth, such as APGAR scores (colour, heart rate, reflexes, muscle tone, and respiration), birth weight and temperature. Neonatal behaviour is also monitored, as conditions such as extreme restlessness or feeding difficulties may also be symptomatic of neonatal pathology. Finally, it can also include indicators recorded from the mother such as blood pH

and oxygen levels, which can indicate some form of abnormal condition or distress prior to or during the birth itself.

Clinical assessment of a neonate is obviously a large topic and an entire specialisation of medicine and will not be covered in detail here. These measures are all integrated by clinical teams when attempting to diagnose and treat neonatal illness. This clinical information can only give an indication of potential pathological presence and allows for only a limited interpretation of the causes and potential long-term effects. However, the brain is difficult to clinically assess directly without imaging techniques. When these symptoms are found and judged to be sufficiently important and relevant to the brain, a request can be made for some form of clinical imaging. This allows medical professionals to directly observe the neonatal brain in vivo and access otherwise unobtainable information about its appearance and condition.

1.4.2 Prematurity

The term period is defined as between 37- and 42-weeks GA and is considered the range in which typical birth occurs. Prematurity is defined as any birth occurring before the lower bound of this prescribed range, with its degree of severity increasing as the GA at birth decreases. Prematurity has been shown to have a strong influence on the neonatal brain and its development, with the likelihood of adverse effects on developmental trajectories and outcomes increasing with degree of prematurity (Dyet et al., 2006).

Higher incidence rates and increasing levels of severity in haemorrhage, specifically germinal matrix haemorrhage (GLH), periventricular leukomalacia (PVL) and punctate white matter lesions (PWML) have all been associated with increasing prematurity in neonates. Further, certain pathologies, such as PVL, can be almost exclusively observed in preterm neonates (Counsell et al., 2003; Volpe, 2009). Structural deformations are also more likely, with the

premature neonatal brain having smaller total and regional tissue brain volumes, and larger ventricular space compared to term neonates (Boardman et al., 2006; Inder et al., 2003; B. S. Peterson et al., 2000).

1.4.3 Haemorrhages, Ischemia and Infarcts

Haemorrhages are one of the more common types of neonatal brain injury and while seen predominantly in preterm neonates, may also be present in term born babies with perinatal complications (Lekic et al., 2015). A haemorrhage is defined as when a blood vessel has become damaged causing bleeding into the surrounding brain tissue. There is great variety in the extent, cause and effect of haemorrhage and are typically described based on their location and cause and their location influences how they present on imaging.

There are several classifications of haemorrhage that are specific to or commonly associated with neonates. A germinal matrix haemorrhage (GLH) occurs as a result of damage to the highly vascularised region of the same name located around the lateral ventricles (Luo et al., 2019). These vary from small, singular bleeds adjacent to the ventricles to more extensive damage into the surrounding tissue. Inter ventricular haemorrhage (IVH) is observed when either an external haemorrhage (frequently GLH) infringes into the ventricle or from damage to the choroid plexus (Ballabh, 2010). IVH usually results in either blood emanating from the injured area or pooling in the inferior or posterior of the ventricles depending on subject orientation. Cerebellar haemorrhages (CH) are often small and globular in nature but can be larger and more irregular in severe cases (Hortensius et al., 2018).

Another potential injury seen in neonates, which may arise due to haemorrhage, is stroke, which occurs due to disruption of blood flow to an area of the brain (Lynch et al., 2002; M. A. Rutherford et al., 2012). Stroke is usually caused by ischemia, wherein a blood vessel of the brain's vascular system is occluded rather than ruptured (Runchey & McGee, 2010). Alongside

the location of the injury itself, stroke is often described as being downstream of some initial ischemic or haemorrhagic injury. If this stroke damage becomes sufficiently severe, the tissue can necrotise, causing blood to flow into the injured area resulting in an injury known as an infarction (V. Miller, 2000). In a similar manner, extensive cases of GLH may develop into a more damaging injury called haemorrhagic parenchymal infarct (HPI) (Bassan et al., 2006; De Vries et al., 2001). HPI is a more specific and severe type of infarction that affects the more critical functional tissue. As such, it is considered more clinically impactful and neonates that suffer HPI are much more likely to have negative or adverse long-term outcome.

As can be seen in several of the injuries, damage that results in necrosis of the tissue are often one of the more extreme results. A total ischemic or haemorrhagic insult can result in irreparable damage and tissue death. However, this can occur on a much larger scale, affecting multiple areas throughout the brain. In the most extreme cases this can result in total, or near total, hemispheric or cerebral devastation, where vast areas of the brain are damaged beyond recovery and completely atrophy (Shankaran et al., 2005).

1.4.4 White Matter Injury

Colloquially, white matter injury has been used as an overarching term to represent injury located in or having some effect on white matter tissue. However, in terms of a specific injury, it refers to damage to the oligodendrocytes surrounding a neuronal axon, resulting in a loss of myelin (Back, 2017; Khwaja & Volpe, 2007). More loosely, it can also refer to injuries that result in the destruction of the white matter itself via direct damage or trauma to the axon or as a downstream injury due to damage elsewhere. Genetic and congenital conditions along with prematurity have been shown to contribute towards both the incidence rate and level of white matter injury severity (A. M. Li et al., 2009).

One of the more common types of neonatal white matter injury is a pathology known as punctate white matter lesions (PWML). These are small focal lesions frequently located in the periventricular white matter (Ramenghi et al., 2007). The origin of PWML is usually a small ischemic or haemorrhagic insult which, depending on time of imaging, may be visible instead and results in disruption of the myelin and potentially necrosis (Hayman et al., 2019). PWMLs can range in severity from small singular cases to, in more severe injuries, multiple florid lesions with an almost conjoined appearance running parallel to the lateral ventricles (Kersbergen et al., 2014).

A diffuse white matter lesion refers to an injury where large areas of myelin have been damaged (Back, 2017). Although affecting a larger area than the more focal PWML, in many instances the loss of myelin is less extreme and the injury itself less severe. This is because the diffuse lesions are associated more closely with loss of oligodendrocyte progenitors rather than damage to already myelinated oligodendrocytes followed by necrosis.

Periventricular leukomalacia (PVL) occurs when the white matter surrounding the lateral ventricles suffers an intense ischemic insult leading to necrotisation and complete atrophy of the tissue (Volpe & Zipurksy, 2001). After the surrounding, less- or un-effected tissue has recovered there will remain a fluid (CSF) filled cavity. Prior to, during and just after initial recover, the surrounding tissue may also exhibit diffuse white matter damage due to gliosis, (Folkerth, 2006).

1.4.5 Grey Matter Injury

As grey matter areas (cortical surface, basal ganglia and thalamus) play a much greater role in cognition and function, lesions in these areas can have a larger effect on development and long-term outcome. Grey matter areas have a higher oxygen requirement when compared to white matter, consuming 93% of the cerebral oxygen supply (Mercadante & Tadi, 2022). This greater

reliance on cerebral blood flow and oxygen supply renders them highly susceptible to interruptions, with damage often resulting in a rapid increase in injury severity (Back, 2014; M. Rutherford, Malamateniou, et al., 2010). Injury in these areas can frequently be attributed to ischemic or haemorrhagic damage to the relevant cerebral arteries (Purves et al., 2001; Tribolet, 2010; Volpe, 2009). The severity of the damage is dependent on several factors including the initial cause, the duration and extent of the injury. In many instances, damage to the grey matter quickly leads to gliosis and neurone loss.

Within the deep grey matter, injuries can range from mild, small lesions which may recover over time and have unknown clinical relevance, to more severe and extensive lesions, which lead to atrophy and loss of a substantial amount of tissue (Boardman et al., 2006; M. Rutherford, Biarge, et al., 2010; Volpe, 2012). There is often a corresponding reduction in deep grey matter volume, with the basal ganglia and thalamus appearing more ragged and containing CSF filled holes. Deep grey matter injury can result in a number of long-term disabilities including cerebral palsy, cognitive impairment, developmental delay and behavioural issues (Martinez-Biarge et al., 2010; Reid et al., 2015; Trivedi et al., 2017).

Damage to the cortical surface is similar to that seen in deep grey matter injury, a hypoxic or ischemic insult resulting in gliosis and cell death. The causes are also similar, frequently the result of some form of disruption to the cerebral arteries, although to a different set to those that cause deep grey matter injury (Purves et al., 2001). These injuries give the afflicted gyri a ragged appearance, an overall reduction in cortical surface volume, larger extracortical space and an increase in total CSF volume. The effect on long term outcome is dependent on both the severity of the damage and the region the injury is located in, such as the motor or somatosensory cortices.

1.4.6 Cystic Lesions

Cystic lesions are another potential pathology found in neonates. A cyst is defined as a fluid (CSF) filled cavity separated from the surrounding tissue by an epithelial cell lining, a feature which the closely related pseudocysts do not share. While cysts can appear singularly, they more often occur bilaterally in the brain, although this is strongly influenced by both their location and initial cause. While cysts and pseudocysts can develop on their own, they are frequently found at the site of previous injury, for example the severe focal lesions which lead to PVL often result in an associated cystic formation (CPVL) (Volpe & Zipurksy, 2001).

One common type of cyst can be found anterior (and potentially superior) to the anterior horns and anterior and inferior to the temporal horns of the lateral ventricles, separated by a thin epithelial wall (Chang et al., 2006). Subependymal pseudocysts are found in the caudothalamic groove or within the wall of the caudate nucleus. These pseudocysts can be formed from an earlier pathology, such as GLH, or via developmental abnormality and may slowly disappear over time (Bats et al., 2002). Subarachnoid cysts occur when CSF penetrates the arachnoid membrane filling the subarachnoid space (Pradilla & Jallo, 2007). This can then extend into the intercranial space and can potentially be found in any part of the central nervous system.

1.4.7 Brain Structural Abnormalities and malformations

Structural abnormalities can be the result of some other initial neonatal or maternal injury, such as enlarged ventricles, or a congenital or genetic condition like lissencephaly. Large amounts of extracortical CSF would suggest that the brain is not of sufficient size to fill the intercranial space and have lower than expected total and regional tissue volumes. This could be the result of either under-development or damage followed by atrophy. Another potential condition is hydrocephalus where one or both ventricles become enlarged due to some factor preventing drainage of CSF from the interventricular space. It can range from slight enlargement of the ventricles to, in the most extreme cases, ventriculomegaly (McAllister, 2012).

1.4.8 Hypoxic Ischemic Encephalopathy

While detection of individual brain injuries is challenging in and of itself, certain patterns of injury can often be associated with or are representative of a larger multi-system pathological condition. For example, specific patterns of white matter lesions are often seen in patients suffering from multiple sclerosis (Granberg et al., 2017). Similarly in the neonatal setting, hypoxic ischemic encephalopathy (HIE) can be represented as lesions in the cortical grey matter, lesions in the periventricular white matter, lesions in the basal ganglia or a combination of the three (M. Rutherford, Malamateniou, et al., 2010). Recognising these indicator injury patterns is a key factor when trying to diagnose a specific pathology.

1.4.9 Summary

Several points need reemphasizing when considering neonatal brain injury. First is the great diversity in the type of lesions which may be present within the neonatal neurological setting. Even within a single type of pathology, there is huge variation in location, size, shape, extent, cause, severity and effect on development and long-term outcome. Even in the event that a single pathology has a well-documented clinical profile, outlier cases of that pathology can still exist. These issues are further confounded by the developmental processes occurring during the neonatal period. These not only affect appearance of healthy tissue but may also alter pathological appearance. A final point to consider is that this has not been an exhaustive list of pathologies, only those closely associated with neonates. Others, such as tumours and specific abnormal structure formations like polymicrogyria or lissencephaly have not been covered and many more exist.

1.5 Methods to Non-invasively Image the Brain

While an initial diagnosis based purely on clinical scores and direct patient observation is possible, it can be greatly assisted by an in vivo examination of areas frequently associated with these altered states. Within a hospital setting, there are multiple neuroimaging techniques

available which can be used for this purpose. These include ultrasound, magnetic resonance imaging (MRI), computed tomography (CT) and positron emission tomography (PET), although the latter two are unsuitable for neonatal imaging due to their radioactive components.

MRI produces high resolution, high signal to noise ratio three dimensional images with excellent contrast between tissue types. Its primary limiting factor is its sensitivity to metal and the limited availability of space within the imaging chamber, which may limit its use in critical cases where constant access to specialist equipment is required. In comparison, the ubiquity and ease of use is ultrasound's primary benefit, hence its regular use in maternal monitoring during pregnancy. This is offset by its poorer image quality compared to MRI and the difficulty in gaining a complete representation of internal structure as acquisition is only two dimensional.

1.5.1 Clinical Neuroimaging using MRI in Neonates

When a neonate is referred for a MRI scan, a number of different imaging modalities are acquired. For a standard clinical acquisition cycle, this will include a T1 weighted image (T1w), a T2 weighted image (T2w) and some form of diffusion imaging. Image modalities highlight normal and pathological tissue in different ways, each of which will contain complementary information that aids the diagnostic efforts (M. Rutherford, Biarge, et al., 2010).

Imaging in neonates contains additional challenges when compared to adults. Good quality MRI acquisition is highly reliant on the individual being scanned remaining as motionless as possible, a task which is understandably difficult in neonates (M. Rutherford, Biarge, et al., 2010; Zaitsev et al., 2015). This is frequently accomplished by either timing scans to intersect with periods of sleep or via sedation of the neonate, the second of which can potentially add additional risks to clinically unstable neonates. There is also the previously mentioned difficulty in accommodating neonates who require constant critical care, as much of the

equipment contain metal components that negatively impact the MRI, pose a potential hazard or are bulky and may be difficult for the scanner to accommodate. In many hospitals that focus on treatment of neonates, specialist equipment is developed that compensates for these difficulties.

Once the images have been acquired, there are further challenges associated with their review and the assessment of a neonate's condition. As noted earlier in section 1.3 there are several developmental processes underway at this time, each of which has an effect on the acquired neonatal images. The neonatal brain is smaller than that of an adult, an issue which becomes more pronounced at younger ages, but MRI images are still acquired at the same resolution. In clinical images, this is 1.4x1.4x4mm to 1mm isotropic in T1w and T2w depending on site and scanner, and 2mm isotropic in diffusion. Alongside this increase in volume over time, is the increase in cortical surface complexity attributable to the ongoing gyrification process. Finally, there is the effect of the ongoing myelination of the white matter tracts, which has a more complicated effect on the MRI images.

Intensity values within an MRI image are relative, with certain tissue types appearing lighter or darker depending on their composition. The major challenge in neonatal imaging is that unmyelinated white matter has an opposing intensity profile compared to myelinated white matter (Barkovich, 2000). This effects not only the white matter intensity values but the grey matter values as well. Not only will white matter intensities now appear inverse to their adult values when compared to the grey matter, either lighter or darker depending on modality, but they will also shift the grey matter intensities in the opposing direction as well (Branson, 2013). The final confounding factor to this challenge is that the myelination process is still ongoing during this period. This results in a gradual change of this unmyelinated image profile to one more closely resembling that seen in adults, a process completed by the first year of life.

1.5.2 T1 Weighted Imaging

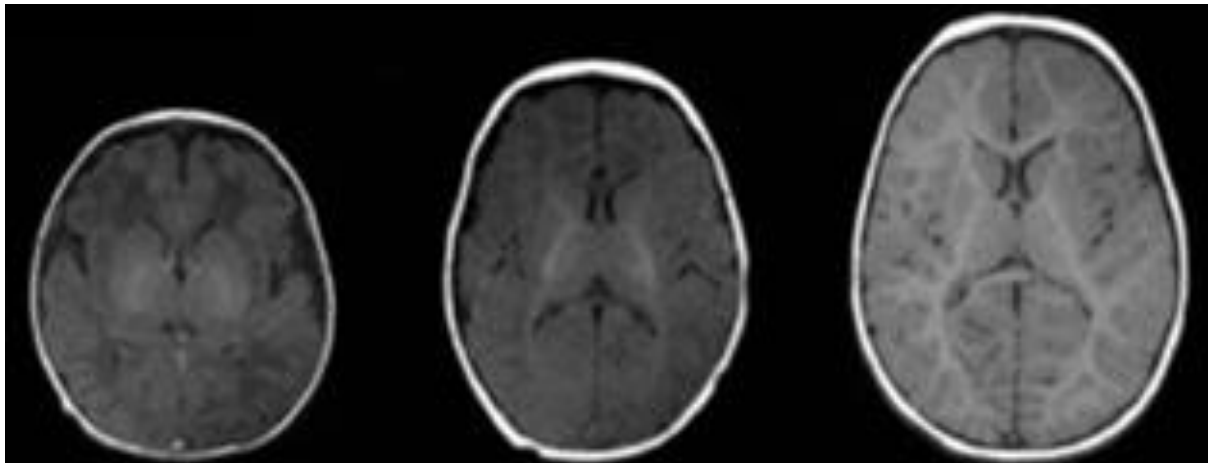


Figure 1.8 T1w images at different ages. From left to right, T1w images of neonates at ages 2 weeks post menstrual age, 3 months and 1 year. Areas of white matter appear dark in the first image but grow progressively lighter as the neonate ages. Images reprinted from Seminars in Paediatric Neurology volume 33, J. H. Miller et al, Neonatal Neuroimaging, Copyright (2020), with permission from Elsevier.

In T1w neonatal acquisitions, myelinated white matter appears as areas of hyperintensity and will usually be the brightest part of an image, the exception being certain lesions and the Pituitary Gland. This is followed by, in descending order of intensity, areas of grey matter, unmyelinated white matter and finally the CSF, which appears as areas of extreme hypointensity. Examples of the different T1w images can be seen in Fig 1.10, showing the changes in intensity which can be observed during the neonatal period (J. H. Miller et al., 2020). These show the change in white matter intensity over time, becoming brighter as myelination progresses, and the effect on grey matter intensity which becomes darker due to the relative intensity change. A common clinical adaption of this is a magnetization-prepared rapid gradient-echo (MPRAGE) image which has the effect of removing the residual CSF intensity from the image (Mugler & Brookeman, 1990).

1.5.2 T2 Weighted Imaging

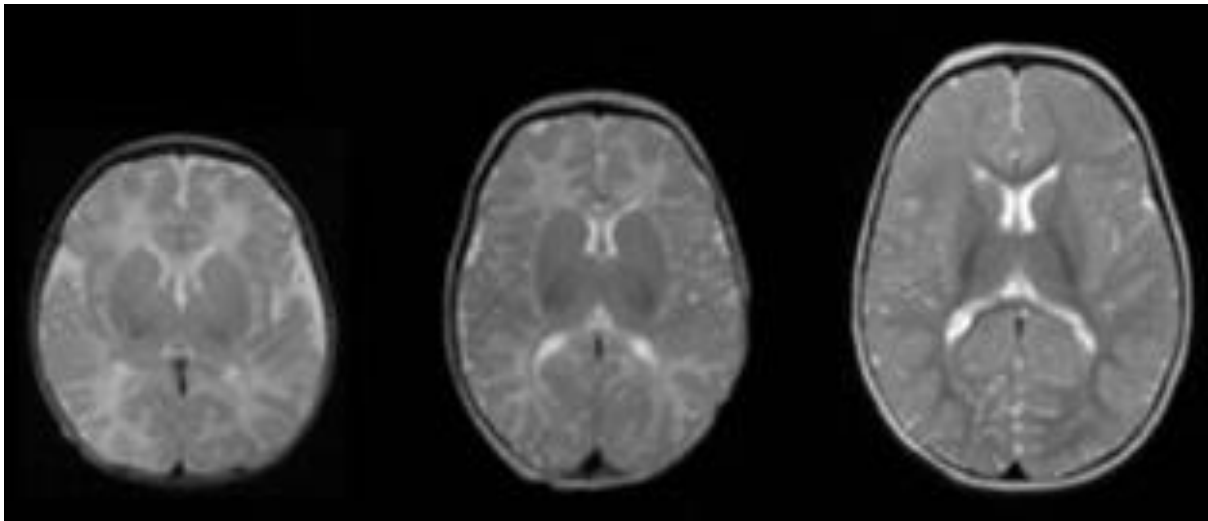


Figure 1.9 T2w images at different ages. From left to right, T2w images of neonates at ages 2 weeks post menstrual age, 3 months and 1 year. Note that areas of white matter appear bright in the first image but grow progressively darken as the neonate ages and that, in comparison to T1w images the CSF is the brightest component of the image. Images reprinted from Seminars in Paediatric Neurology volume 33, J. H. Miller et al, Neonatal Neuroimaging, Copyright (2020), with permission from Elsevier.

For neonatal T2w images, the highest intensity values can usually be found in the CSF, Fig 1.11. This is followed in descending order by unmyelinated white matter, grey matter then areas of myelinated white matter. In a similar manner to T1w's MPRAGE there is an additional acquisition modality called Fluid-attenuated inversion recovery (FLAIR) which is used to suppress CSF (De Coene et al., 1992). While a MPRAGE image does not appear drastically different from a T1 image, FLAIR has a much more noticeable effect as it removes the normally bright CSF from the image allowing for better differentiation between the remaining tissue.

1.5.3 Diffusion tensor Imaging

Diffusion tensor imaging (DTI) models the diffusivity, ability to move freely, of water molecules within some medium, in the context of this project the neonatal brain tissue and CSF

(Elster & Jonathan H. Burdette, 1994). In pure water, the diffusion of any given water molecule will be random as it may move freely in any direction, known as isotropic diffusion, and will have a high diffusion coefficient. If diffusion is restricted, for example by impurities within the medium such as macromolecules or boundaries tissue membranes, it results in more limited movement within certain directions (anisotropic diffusion) and will have a lower diffusion coefficient.

A diffusion image acquisition begins with what is known as a b_0 image, a T2w image with no additional diffusion features that is used for calculation of later maps such as apparent diffusion coefficient (ADC). For the diffusion acquisition, two additional pulses are needed, the first knocks molecules out of phase alignment while the second realigns them. However, this realignment will only take place in stationary molecules (tissue or highly restricted water), molecules which can move freely will not be exposed to the second pulse and will continue to de-phase. This allows the diffusivity to be measured based on if the molecules within a voxel have recovered their phase alignment or not. Diffusion sequences are acquired in a minimum of three directions, although frequently more are used (e.g., 16 and 32 directions), where each is perpendicular to the other two. This allows creation of a tensor that measures the level of diffusivity in each of the sampled directions.

While other visualisations are possible, within the clinical setting DTI images are often displayed as ADC, sometimes called mean diffusivity. Areas containing free moving water, such as CSF, will have high diffusivity levels and appear as areas of high intensity. These voxel's intensity levels will gradually decrease as the level of water diffusivity becomes lower. Development of paediatric ADC atlases by Ou et al., 2017 found there to be a general reduction in ADC values over time. The highest within tissue ADC values were found in areas of unmyelinated white matter, this was followed by myelinated white matter and finally grey matter. The advantage of this method, unlike T1 and T2 weighted images, is that the values are

quantitative and can be directly compared between infants where the acquisition approach is identical.

1.5.4 Pathological Appearance

In many instances, injuries within an image appear as areas of hypo/hyperintensity, e.g. injuries such as lesions or haemorrhages appearing as bright/dark areas depending on modality (Dorner et al., 2018; Hinojosa-Rodríguez et al., 2017; O’Muircheartaigh et al., 2020; M. Rutherford et al., 2004). However, certain abnormalities, such as enlarged ventricles and cysts, do not have these obvious hypo/hyper intense values. As such detection will rely on additional contextual information such as differences in expected intensity values compared with those seen in typical structure. Alternatively, detection may be possible via identification of neonates with gross morphological differences, although this does not inform on the location and extent of an injury (Dimitrova, Arulkumaran, et al., 2021; Tarui et al., 2020).

White matter lesions appear as areas of hypo-intensity in T2w images and areas of hyperintensity in T1w images compared to areas of healthy white matter (O’Muircheartaigh et al., 2014). Haemorrhage appears as areas of hyperintensity in T1w and hypo-intensity in T2w images, although to a greater extent than that seen in lesions (Dorner et al., 2018). Cystic tissue has an opposing T1w and T2w intensity profiles, with CSF intensity values being present in areas that should contain white or grey matter (Counsell et al., 2003). Damage to the cortical surface can be seen as areas of higher-than-expected intensity on T1w images (M. Rutherford, Biarge, et al., 2010). Damage to the basal ganglia and thalamus can be highly variable with both high and low intensity areas, sometimes described as a ‘marbling’ effect, compared to normal deep grey matter values. Injury seen in ADC maps has a highly variable appearance as damage can appear as areas of either higher or lower levels of diffusion depending on the time and level of injury. Examples of different pathologies are shown in figure 1.12.

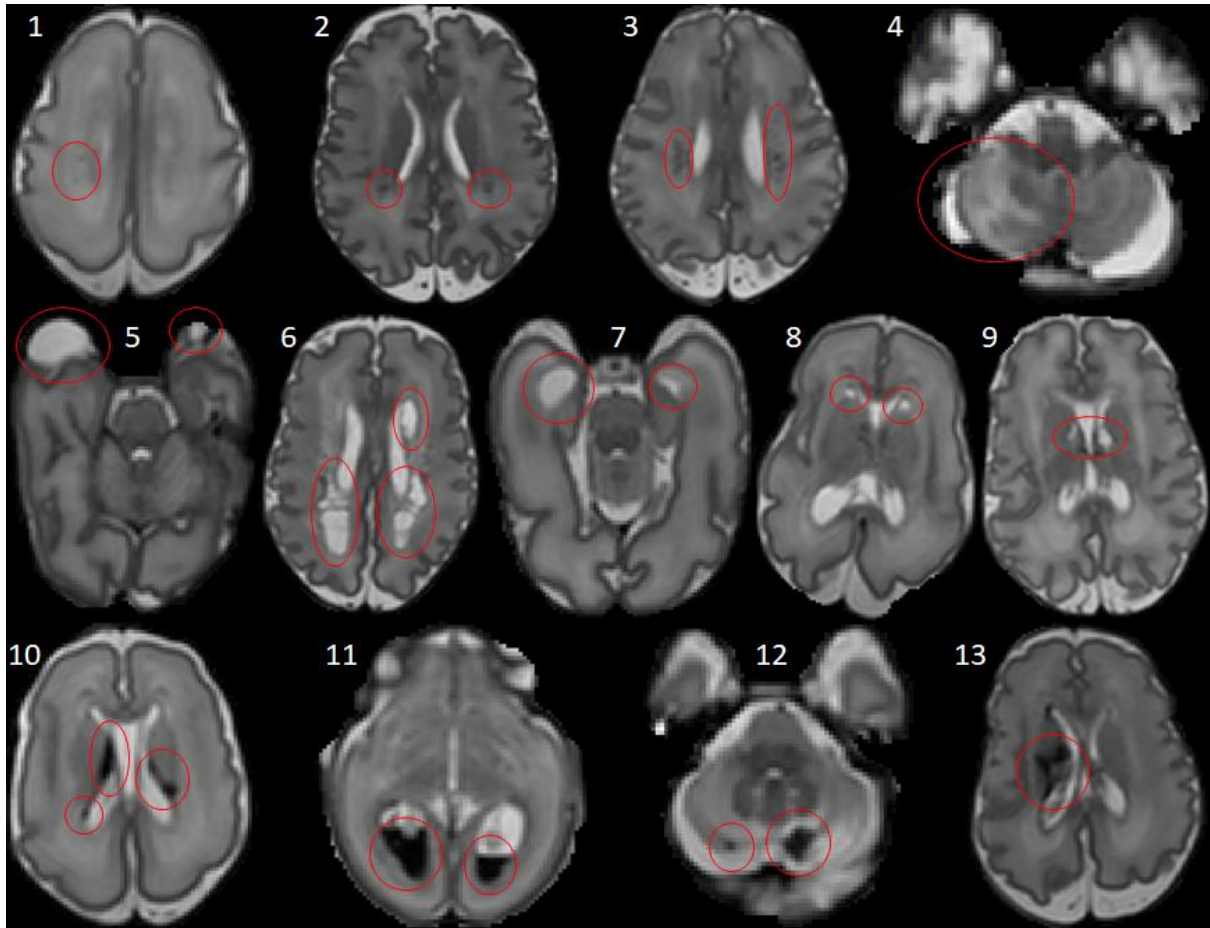


Figure 1.10 Images of neonatal pathology. Different injury patterns that can be found within pathological neonates, the red circles outline the injured areas. Images 1, 2 and 3 show very subtle, more obvious, and florid punctate white matter lesions, while image 4 shows an ischemic injury in the cerebellum. Images 5-9 show cystic tissue, specifically sub-arachnoid cysts (5), cystic periventricular leukomalacia (6), cysts on the temporal horn of the lateral ventricles (7), pseudocysts on the anterior horns of the lateral ventricles (8) and subependymal cysts (9). Images 10-13 show haemorrhagic injuries including germinal matrix haemorrhage (10), intraventricular haemorrhage (11), small and large cerebellar haemorrhages (12) and haemorrhagic parenchymal infarct (13).

1.6 The Need for Computer Assisted Medical Detection

An additional challenge often associated with medical image review is one of inter and intra-rater variability (Bogner, 2018). While the core areas of injury can often be consistently highlighted, there can be disagreement over the exact extent of an injury. This can lead to

inconsistency when estimating what tissue should be considered part of the injury, depending on how liberal or conservative a reviewer is being. This can also be true for diagnosis made by a single review over multiple images or even in the same image at different times. Experience can often have a large effect, as reviewers become more familiar with the medical (or personal) consensus on what constitutes lesioned tissue.

Unfortunately, there can be additional effectors over which reviewers have little control. Clinical radiologists with additional specialisation in the paediatric and neonatal settings are rare with many hospitals having to refer these cases to institutions with the appropriate personnel. This can lead to long hours for the limited number of radiologists available, and, potentially, limits on the review time available for a single case. This can have strong negative effects on the review process. With the current rate of staff turnover and limited uptake in training of new radiological staff, it is predicted that only 44% of the needed radiological personnel will be available by 2025 (*Clinical Radiology UK Workforce Census 2020 Report*, 2021).

As stated earlier in this chapter, reviewing neonatal MRI is a highly specialised task and while the skills and experience needed to carry out proper assessment may be available in specialised clinical settings, this may not necessarily be the case. This situation is likely to be further exasperated, likely to an greater extent than general clinical staff. It is for this reason the computer aided detection could prove so critical in the future. By allowing machine learning algorithms to handle certain tasks, such as general triage of MRI images, it would allow radiological staff to focus on the much more difficult tasks have assessing injury level and prognostic evaluation.

As yet human assessment still outstrips computer controlled medical diagnosis, a situation that is unlikely to change in the near future. However, these still imperfect assessments could offer

alternative interpretations, potentially acting as semi-informed alternative opinions or assisting newer clinicians when more experienced staff are unavailable. Final decision on the radiological assessment remains outside the purview of current machine learning efficacy levels, leaving it firmly in the hands of the radiologist performing the review. Using computer aided diagnosis to assist this process and potentially making it more robust is within the realm of possibilities, a situation which this thesis hopes to demonstrate.

1.7 Chapter Summary

The primary goal of this thesis is to create a generalised abnormality detection method for use in neonatal MRI. This method must be capable of detecting the wide range of pathologies which can be present during the perinatal period. To achieve this, the model must be applicable to the multiple MRI modalities commonly acquired in clinical MRI scans; T1w, T2w and ADC image. Additionally, it must be able to make these distinctions while accommodating the changing appearance caused by the various developmental processes (growth, myelination and gyrification) found during this period (28-55 weeks GA). Ideally, this method should be able to both identify neonates within a cohort who have some form of abnormality present and, within a given neonate's individual MRI, identify areas that are likely to contain these abnormalities. Finally, this work hopes to use these individual measures of abnormality in order to estimate the likelihood of a neonate having either a normal or disrupted long-term outcome.

Chapter 2. Machine Learning

2.1 Introduction and terminology

The primary focus of this thesis is to investigate the use of computer aided diagnosis (CAD) for detection of injured tissue, pathology and structural abnormality within neonatal MRI. Before this can be undertaken, an understanding of the different machine learning (ML) methods is necessary. This chapter will act an introduction to this field, how to categorise the different ML techniques and strategies for validating their success and accuracy. It will also cover give an overview of common ML algorithms, how they are frequently utilised and their advantages and disadvantages. This will be followed by some examples of how specific ML techniques have already been used within the neuroimaging field.

At the highest level, ML algorithms work with some input data and optimise components of a given machine learning method in order to help perform tasks (the output). The form of the data can be continuous (e.g., height and weight) or discrete (e.g., binary or categorical values, such as healthy and diseased tissue or male and female).

Data used in ML can often include many features (multivariable or multidimensional), some of which can be duplicates or superfluous for the purpose of output prediction. A frequent preparatory step used prior to training an ML method is a process known as feature selection; the reduction of input variables to those that explain most of the variance in the data or which keep the most informative features. Methods such as independent and principal component analysis (ICA and PCA) can be used to reduce model complexity. Another common technique is feature expansion; the inclusion of addition complex input variables such as combinations of or polynomials of base input features. As an example, prior knowledge that perinatal and postnatal brain growth follows a sigmoid shape can be incorporated into an ML method, allowing a more parabolic functional mapping which may better fit the data.

Model generation or construction refers to the process of selecting the components or architecture of a given model which will be described in detail later. This can also encompass picking which ML algorithms are implemented. The application of an algorithm to the data can be described in multiple ways in the literature including fitting, optimising or generating a model or just modelling. Optimisation or fitting is the process of selecting the optimal values for any parameters or hyper-parameters associated with a given ML algorithm for a specific task. The term modelling is also sometimes said to apply to the entire process, including, some form of testing to assess the fitness of the proposed solution (validation).

In a ML context, a parameter can be described as a weighting applied to a single, combination of or derivation of input features/variables which map them to some output variable. In contrast, hyper-parameters affect model architecture or parameter optimisation process and are specific to a given ML method. In practice this can include choices such as the optimiser, improving model accuracy or speed of optimisation, or regularisation method, down-weighting the importance of extreme values. These hyper-parameters will need to either be manually selected and tested or may have their own optimisation process.

2.1.1 Model Validation

To assess model performance, available data will often be divided into training and testing sets and, often, a third external validation set to allow for optimisation and assessment of model efficacy. Testing the generalisability of a model is performed via a method called cross-validation, of which there are several types, a basic implementation of which is the train/test split. The training set is used for the optimisation process which, in a simple form, may consist of 80% of the total available data and the trained, optimised model is then applied to the remaining 20% testing data to estimate how well it performs on unseen samples. A common variation on this is 5-fold cross-validation, which repeatedly holds out a different 20% each time.

This can be generalised to another method called K-folds. K-fold cross-validation splits the data into K equally sized sub-divisions, each of which will have their own model trained on K-1 of the sets and tested on the remaining 1, with 5-fold and 10-fold being popular choices for K. A related method is stratified k-fold cross-validation which keeps the number of samples of each output class proportionally the same, that is if there is an imbalance in the classes this will be preserved in the training and testing sets. A similar method often used for small datasets is leave-one-out or leave-p-out cross-validation, which iteratively holds out either a single or p samples for testing while training on all other samples. The final method that sees common use is Monte-Carlo cross-validation, which builds k models, where each model will randomly subsample the data to select training and testing sets while ensuring the size of these sets is different for each model.

These cross-validation methods can also be nested. The dataset is split using a method such as k-fold or leave p-out as usual, forming the outer loop. However, within each training set, an additional cross-validation is performed, the inner loop. The performance of each model within the inner loop is compared and used to select the best parameters and hyperparameters. These are used to create a model which is applied to the held-out testing set of the outer loop to assess overall ability. The performance of all models applied to the outer loop is then averaged to get a robust interpretation of the overall performance.

One issue affecting the choice of cross-validation method is the sample size available. Many ML methods, and cross-validation techniques, are frequently applied to large datasets, ranging from thousands to millions of samples. Within the neuroimaging field, dataset sizes are much smaller, frequently being composed of less than 100 samples. As noted in Varoquaux, 2018, these smaller sample sizes result in inaccurate estimation of performance. As the test set size becomes smaller it also becomes more likely that accuracy errors will be underestimated. This is obviously accentuated by use of techniques that attempt to leverage as much information as

possible during training, such as leave-one-out cross-validation, which reduce the testing set to minimal levels. As such, use of 5, 4, 3 or even 2-fold cross-validation or leave-p-out with large values of p may be preferable to maximise the robustness of accuracy estimation, even at the cost of a reduction in maximal accuracy score.

2.1.2 Machine Learning Approach Strategies

A complete learning strategy can include a single or multiple ML approaches. Each technique can belong to one or more broad categories depending on what data they use, how they model this data and the format their output takes. This section will give a broad overview of these categories, how they are used and some common techniques. I will do this in terms of splits of families of algorithms and the tasks they perform.

The first categorical split is between supervised and unsupervised methods. Supervised methods are those in which both input and output variables are known, allowing a calculation of some functional mapping between them. In comparison, unsupervised techniques initially make use of only the input variables and typically generate reduced rank representations. Unsupervised techniques are often used as an initial step to identify key features or groupings within the data before some secondary technique.

An example of an unsupervised method is clustering algorithms which attempts to group individual data points into discrete (latent or unknown) categories based on the input variables. Others include similarity dimensionality reduction techniques such as PCA and manifold learning which try to reduce the number of active features by projecting the datapoints into lower dimensional space while preserving their differentiability. Example of a supervised ML methods include linear and logistic regression, support vector machines (SVMs) and neural networks all of which will be covered in more detail below.

A second grouping can be made based on whether a model is parametric or non-parametric. Parametric methods model their outputs based on some mathematical combination of the input features, such that they can be directly mapped to the output variable. In comparison nonparametric models are not constrained by a functional form and instead make predictions based on similarity between the training samples and the testing samples. They assume that unknown outputs will have classes or values like those known samples with similar input variables. Exactly how a point is considered similar is often the result of an optimised or specified strategy or hyper-parameter of the given ML algorithm. Common examples of parametric models are multiple linear regression and neural networks, which generating extremely complicated function mappings, but do so based on the initial input variables. Non-parametric models include SVMs, which project data into a higher dimensional space to help separate different classes, and Gaussian processes, which make prediction based on the similarity between the prediction target and a set of known trainings samples.

Another distinction is between linear and non-linear models. Linear models assume that the relationship between the input features and output variable is strictly linear allowing predictions to be made by a direct mapping if parametric. This assumption holds for features generated via manual feature selection, e.g., using the square of an initial variable allows the creation of curved polynomial function but is still considered linear. Non-parametric models, such as linear SVM and k-means clustering, can also be linear, although in this instance it refers to the separation boundary applied to the samples. In comparison, non-linear models can be considered any model which does not have the previous constraints. Some examples of non-linear models include neural networks, which are parametric and who's many layers allow automatic feature selection, and SVMs, which are non-parametric and project the input variables into an expanded feature space. Examples of linear models include linear and logistic regression.

Finally, there are regression and classification algorithms working with continuous and discrete outputs respectively. Regression algorithms assume that the output variable will be some continuous value and attempt to predict what this value is given the input variables. Classification algorithms are used when the output variable will have a discrete set of possible classes and attempt to separate each sample into one of these. Regression models output a single value, occasionally with accompanied by a confidence measure that allows flexibility around this exact answer. Depending on the classification method, a sample's output may be single class or member of multiple classes. While this may be preferable in certain tasks where a probabilistic measure is beneficial, some final decision must be made usually by finding the class with the highest membership.

While some machine learning models are either continuous predictions (linear regression) or binary / categorical classification (logistic regression), many models are capable of both depending on their implementation. Indeed, logistic regression can be considered a version of continuous linear regression converted to work as a classification method simply applying a sigmoid function to the output. Examples of regression models include linear regression (supervised, linear, parametric), support vector regression (SVR) (supervised, linear, non-parametric) and Gaussian processes (supervised, non-linear, non-parametric). Classification model examples include neural networks (supervised, nonlinear, parametric) and random forests (supervised, nonlinear, non-parametric). Although as stated earlier, both the neural network and random forest can both be implemented for regression problems. An overview of exactly which groups each method belongs to is shown in Table 2.1.

Table 2.1 Machine learning algorithms. This table outlines many of the machine learning algorithms that see common use in the field of neuroimaging. It also shows the different subdivisions into which each method can be placed.

Method	Supervised/ Unsupervised	Linear/ Non-Linear	Parametric/ Non-Parametric	Regression/ Classification
Linear Regression	Supervised	Linear	Parametric	Regression
Logistic Regression	Supervised	Linear	Parametric	Classification
K Nearest Neighbour	Supervised	Non-Linear	Non-Parametric	Classification
K-Means Clustering	Unsupervised	Linear	Non-Parametric	Classification
Decision-Trees / Random Forest	Supervised	Non-Linear	Non-Parametric	Classification
Naive Bayes	Supervised	Linear	Parametric	Classification
Markov Random Field	Supervised	Linear	Parametric	Classification
Support Vector Machine	Supervised	Both	Non-Parametric	Classification
Neural Networks	Supervised	Non-Linear	Non-Parametric	Both
Gaussian Process Regression	Supervised	Non-Linear	Non-Parametric	Regression

2.1.3 Measure of Performance

Assessing the model performance is as important as the construction itself. This is partly to ensure that, at a bare minimum, the model is outperforming random chance and is useful. For example, when performing binary classification with a balanced dataset a random accuracy would be 50% so scores higher than this would be required. Unbalanced datasets need additional considerations as metrics must account for the level of imbalance. For example, if a

dataset has two classes but one accounts for 75% of the data, a model that simply assigns that class to all datapoints would produce an accuracy of 75%, overestimating its performance.

The type of performance measure used is dependent on the type of task being assessed, with mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE) seeing use in regression problems while classification is based on the labelling of datasets into true and false positives and negatives. These scores not only ensure a minimum level of performance but also for a comparison between models.

2.1.3.1 Assessing Regression Algorithm Performance

Mean squared error (MSE), formula 2.1, and root mean squared error (RMSE), formula 2.2, are two closely related metrics of regression accuracy. These are calculated by squaring the difference between the model prediction outputs and the actual values and averaging them over all samples, these can then be used as is, MSE, or rooted, RMSE. Larger errors are magnified compared to multiple smaller errors resulting in harsher assessment for models which make few large inaccurate predictions compared to multiple small inaccurate prediction. RMSE is preferred over MSE when interpretability is desired. The unit of measurements for MSE is the square of the initial units being measured, while they RMSE returns the error value back to the original unit scale make it easier to compare.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

2.1

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

2.2

MSE and RMSE also see common use within ML algorithms as optimisation functions. In these instances, a model is often optimised by attempting to find the optimal parameters which minimise this error, again preferring functions that produce many small errors compared to functions which produce a few larger errors.

Another common metric used to assess regression algorithm performance is the mean absolute error (MAE), formula 2.3. First the absolute difference between all subjects actual and predicted output is found then summed together and finally divided by the number of samples. The absolute function is needed as there is no square effect converting errors into positive values, a necessary step in order to ensure errors of opposing magnitudes do not cancel each other out. The MAE gives equal weighting to all errors rather than overly punishing larger errors, preventing larger errors from causing excessively poor scores.

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}$$

2.3

When working with these scoring systems, values closer to 0 indicate better algorithmic performance. One thing to be aware of is the scale of the output values as this will alter the scale of the error. An error of 0.05 where outputs are between 0 and 1 is the same as an error of 0.5 for outputs between 0 and 10 and an error of 5 for outputs between 0 and 100. This can make comparing similar methods applied to different datasets more difficult, such as predicting brain age in adults vs brain age in neonates where not only the scale is different but also the range, potentially 28-52 weeks vs 18 to 80 years.

There are conflicting opinions about which of the two branches of error estimation are preferable. Willmott & Matsuura, 2005 favours MAE over (R)MSE as the latter becomes proportionally larger as the error variance increases and therefore an unfair representation of the overall average error. In contrast, Chai & Draxler, 2014 argue that (R)MSE give better

estimation of average error on the assumption that those error values follow a gaussian distribution, although they also mention that a truly robust estimation would use both.

An additional consideration can be made based on the task being attempted. In the situation where general overall performance is preferable, MAE may offer the better solution. If the more extreme prediction errors have no accompanying increase in negative effect, there is little reason to additionally penalise them. Conversely, in situation where negative effects need to be minimised at all costs, such as in safety critical situations like aircraft or nuclear power control, and extreme errors are associated with greater risk, (R)MSE would make more sense, precisely because of its greater emphasis on these errors.

For this project, MAE may have an additional use by providing highly interpretable visualisation of model performance. If applied to a slice from an MRI image, general model performance can be identified as a heatmap, allowing identification of good and poor performing areas. MSE or RMSE would be less suitable for this task as they can be easily skewed by large single sample errors.

2.1.3.2 Assessing Classification Algorithm Performance

Classification is best introduced with binary classifier, for example identifying and separating neonates with some form of pathology (positive) from those without (negative). Building on this, a sample's prediction can occupy one of four states. If a target has been predicted as pathological and is pathological it is considered a true positive (TP) and if a sample has a prediction of nonpathological and was nonpathological it is considered a true negative (TN). Conversely, if a sample is pathological but was predicted to be nonpathological it is a false negative (FN) and when a nonpathological neonate has a pathological prediction, it is a false positive (FP). These four states are often combined into a structure called a confusion matrix as shown in Fig 2.1.

		Actual Condition	
		Pathological	Nonpathological
Predicted Condition	Pathological	True Positive (TP)	False Positive (FP)
	Nonpathological	False Negative (FN)	True Negative (TN)

Figure 2.1 Binary confusion matrix. A confusion matrix shows the number of correctly identified True Positives, False Positives, False Negatives and True Negatives and how they are assigned.

This can be expanded to multi-class representation, for example classification of pathological severity as shown in Fig 2.2. In this instance each level of severity will have its own confusion matrix. This is more intuitively seen by looking at a single class example, such as identifying low severity samples (top matrix). A TP is when a model correctly identifies a low severity sample, a FP is when the model assigns low severity to either medium or high severity samples and a FN is when the model assigns either high or medium severity to a low severity sample. TPs are composed of all remaining samples regardless of their actual level of severity or that assigned by the model. For the low severity confusion matrix, it is irrelevant if medium and high severity samples are correctly identified, only that they were not falsely identified as low severity.

This can be viewed in the confusion matrix itself. TPs are identified at the intersection of the predicted and actual class labels (green), e.g. predicted and actual low severity, predicted and actual medium severity and predicted and actual high severity. These will start at the top left for the first class and proceed diagonally to the bottom right as the classes iterate. FPs (red) are composed of the remaining cells of the same row as the TP, for example the top row for the 1st class, 2nd row for the 2nd class. FNs (yellow) are the remaining cells in the same column as the TP, for example the 1st column for the 1st class and 2nd column for the 2nd row. TNs (blue) are the remaining cells in the confusion matrix.

			Actual Severity		
			Low	Medium	High
Low Severity Matrix	Predicted Severity	Low	True Positive (TP)	False Positive (FP)	False Positive (FP)
		Medium	False Negative (FN)	True Negative (TN)	True Negative (TN)
		High	False Negative (FN)	True Negative (TN)	True Negative (TN)
			Actual Severity		
			Low	Medium	High
Medium Severity Matrix	Predicted Severity	Low	True Negative (TN)	False Negative (FN)	True Negative (TN)
		Medium	False Positive (FP)	True Positive (TP)	False Positive (FP)
		High	True Negative (TN)	False Negative (FN)	True Negative (TN)
			Actual Severity		
			Low	Medium	High
High Severity Matrix	Predicted Severity	Low	True Negative (TN)	True Negative (TN)	False Negative (FN)
		Medium	True Negative (TN)	True Negative (TN)	False Negative (FN)
		High	False Positive (FP)	False Positive (FP)	True Positive (TP)

Figure 2.2 Multi-class confusion matrix. A multi-class confusion matrix is a more general implementation that, nonetheless, makes calculation of True and False Negative and Positive more difficult. This image gives a visual demonstrates of how each can be calculated.

From these elements, a number of success measures can be calculated, each of which represent different information about how well a given algorithm is working. Sensitivity (true positive rate, recall), formula 2.4, is the proportion of pathological neonates that were correctly identified, and specificity (true negative rate), formula 2.5, is the proportion of nonpathological neonates that were correctly identified. A third measure, called precision, formula 2.6, states what proportion of neonates identified as pathological actually had some form of injury and accuracy, formula 2.7, the proportion of all predictions that are correct.

$$\text{sensitivity or recall} = \frac{TP}{TP + FN}$$

2.4

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

2.5

$$precision = \frac{TP}{TP + FP}$$

2.6

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

2.7

Other measures give estimations on how well a set of predictions for a given class overlap with their actual class, a measure seen frequently in segmentation tasks. These include the Jaccard index, formula 2.8, and the closely related dice coefficient (dice overlap, F1 score), formula 2.9.

$$jaccard = \frac{TP}{TP + FP + FN}$$

2.8

$$dice = \frac{2TP}{2TP + FP + FN}$$

2.9

In certain circumstances, such as where a threshold or hyper-parameter can be altered, it is possible to calculate the trade-off between the two metrics. The receiver operating characteristics (ROC) gives a visualisation of a given method's trade-off between the sensitivity and specificity, Fig 2.3 left image. An ideal model has only true positives and true negatives, a point at the top left of the plot, Fig 2.3. As a general rule, as sensitivity (true positive rate) increases, higher percentage of correct positive predictions, the specificity (false positive rate) decreases, lower percentage of correct negative predictions, or vice versa. If a model made all positive predictions it would result in perfect sensitivity but zero specificity while all negative predictions result in the reverse situation.

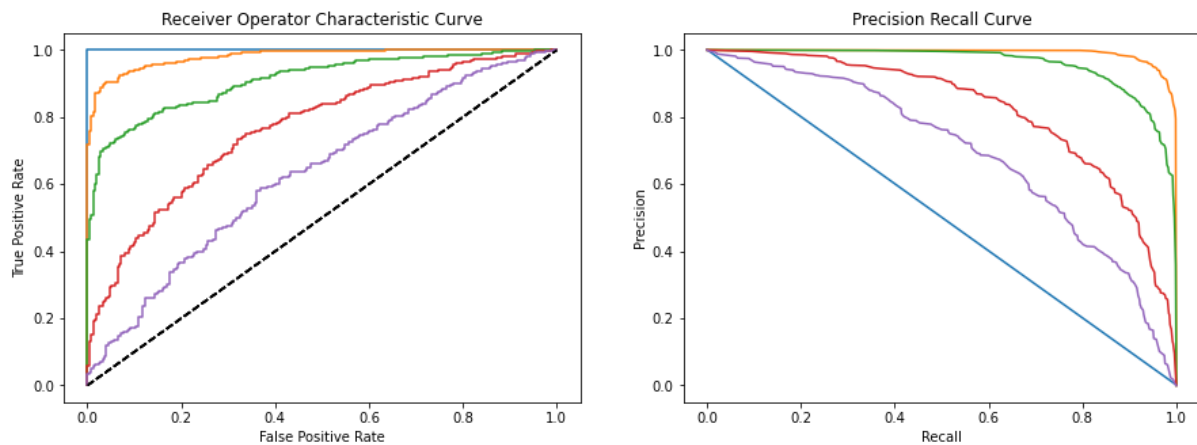


Figure 2.3 Receiver operating characteristic and precision recall curves. The left image demonstrates different levels of performance for the receiver operating characteristic, with the ideal case being the blue line, having perfect sensitivity and specificity (true positive rate and false positive rate), and the purple line being the poorest, showing only just above a linear sensitivity specificity trade off (dotted black line). The right image shows different levels of performance for the precision recall curve, with the orange line having the best performance, almost all positive predictions were true with almost very few false positives. The blue line shows when the situation when the number of true and false positives increases in equal proportion and represent the worst-case scenario.

The precision recall curve (PRC) shows how good the positive predictions of a single class is, Fig 2.3 right image. Ideally there are only true positive predictions, the top rightmost point in the plot, but as the model makes more positive predictions the proportion of these that are true positive is likely to decrease. In this instance, predicting all positive values would result in a lower proportion of true positives to total predicted positives. One of the primary uses of the PRC curve is to assess performance in the case of unbalanced datasets where it gives a better estimate than an ROC curve. This is because ROC gives an estimation of performance of both classes while the PRC focuses more on the less representative class. In the event of an uneven split, e.g., 90/10%, the poor model performance can be seen more readily in an PRC curve compared to what would appear to be reasonable performance on an ROC curve Fig 2.3.

The overall effectiveness of the model can then be summarised by calculating the area under the curve (AUC) of one or both of these curves. This gives an interpretable indication of how well a given method is performing, with values closer to 1 indicating a more successful model.

When looking at these metrics, it is important to be aware that they have the potential to over- or underestimate performance. Frequently this can be attributed to some imbalances between the number of samples in each class. In an ideal world, all classes that a given ML algorithm is attempting to separate would have a roughly equal distribution of samples in both the training and testing sets. However, the situation can arise where one class can have significantly more samples than the other. A good example of this imbalance in biomedical imaging is between the relative proportions of healthy and pathological tissue. Even in subjects with relatively large lesions, there is likely to be a much higher proportion of healthy tissue compared with lesioned tissue. This makes it much easier to achieve high specificity scores by simply having a model that minimises the amount of healthy tissue incorrectly classified as pathological.

There can also be task related considerations, where maximising one metric, for example the sensitivity, is preferable. Usually, this comes in the form of setting a minimum value for one measure, then attempting to maximise the value of the second measure. An example of this is in cancer screening, where a minimum high level of sensitivity is required to ensure that as many patients as possible are diagnosed. Only once this is met, do efforts to minimise specificity begin.

When deciding how best to optimise these scores, it is also important to take into account the detectability, incidence, and effect of different pathologies. For the sake of example, imagine two pathologies, one has serious long-term effects but is both difficult to detect and has low incidence rates, the other is much more common, easy to detect but only rarely results in negative long-term outcome. Frequently, a decision must be made about which of these

potential pathologies a model should be optimised for, and there is often more nuance in the incidence, effect and detectability of different pathologies. Then there are real world material constraints that need to be accommodated such as sample availability, time, staff, computational resources and all of their associated costs.

2.1.4 Scaling of Input and Output Data

In its raw form, the data used for ML can often be in widely varying scales both between subjects and between different variables. Regarding the work in this thesis, the demographic information includes ages (gestational age and postmenstrual age) given in weeks and while consistent between neonates is within the unusual range of 28-55 weeks. Additionally, MRI images have intensity values that vary between individuals, even those of the same modality and for the same tissue. Before any of this information can be utilised by a ML algorithm it will require some form of feature scaling to bring them into broad alignment.

Feature scaling commonly refers to one of two possible applications: normalisation and standardisation. Normalisation (Min-Max scaling) involves shifting all values of a single feature by their minimum value, or alternatively the mean, then scaling by their range, bringing them all into the range of 0 to 1, formula 2.10. Standardisation (mean scaling) shifts all values by the feature mean then scales by its standard deviation, formula 2.11. The features new values have units of standard deviation centred around 0.

$$Normalisation = \frac{X - \min X}{\max X - \min X}$$

2.10

$$Standardisation = \frac{X - \mu}{\sigma}$$

2.11

2.2 Commonly used Machine Learning Methods

When looking at how machine learning can be used within neuroimaging, it is necessary to first identify exactly what type of task is being attempted. The first thing to consider is whether it is being applied within a single subject or across multiple subjects. The initial goal of this project is locating subject specific areas of abnormality and can therefore be construed as a within image problem. The second goal is attempting to separate subjects who have some form of pathology, visible in MRI, from those who have a normal appearing brain and as such is an across subject problem.

The choice of which ML method is used to achieve these tasks is also dependent on what information is available and how it can be presented. Within the radiological neuroimaging setting of this project, information is available as demographics, such as ages and sex, and 3-dimensional MRI brain imaging in several modalities. Demographics are routinely used in ML projects and can be accommodated by their inclusion as features into the design matrix of an ML algorithm. Use of the 3D images, as either inputs or outputs, can be more challenging. While application of ML method to imaging data is a common practice, this is usually in a 2D format. The addition of the third dimension will require either refinement of these methods, use of specially developed ML algorithms designed for this task or some form of dimensionality reduction techniques applied to the data.

While it may be possible to use the whole MRI image, an alternative would be to use some substructure which can be extracted from the 3D image. The first of these to be considered would be dividing the 3D image into 2D slices along one of the three planes, commonly identified in neuroimaging as the axial, coronal and sagittal planes. This has the advantage of making many of the existing ML approaches used for traditional imaging problems applicable for this project. The issue is that it may result in the loss of important contextual information contained in the adjoining slices.

Another possibility is that each slice can be further subdivided into individual pixel components. In the context of a 3D image, the term pixel is inaccurate as it is not a simple 2D object but will also have volume, and therefore should be referred to as a volumetric pixel (voxel). Just as each pixel can be indexed by a vertical and horizontal component which describe its location within a 2D image, a voxel can be indexed by a vertical, horizontal and depth component which specify its location within a 3D image. This could allow a given model to focus on a very specific task, a single voxel, reducing its computational complexity. Such methodologies are often referred to as voxel-wise implementations.

As specified in the introduction, a common task within neuroimaging is segmentation, a process usually completed by reviewing an MRI image and manually selecting voxels within a slice that match the segmentation parameters. Indeed, the first goal of this project would ideally produce an output similar to a manual lesion segmentation, some form of 3D map which delineates potentially abnormal areas of a neonate's brain.

The next task of the project will focus on separating images into discrete classes. Usually, this is a simple binary classification between images containing a specific pathology and images comprising healthy controls. In some instances, it can be more comprehensive, such as having separate classes for different pathologies or identification of subsets, such as levels of severity within a pathology. These types of classification can be enhanced by within image identification of features known to be associated with specific medical conditions. For example, using highlighted tissue lesions and malformations, first identified by other ML methods, or use of the injury localisation maps the project hopes to create.

Machine learning methods can be divided into separate groups depending on the type of task being performed, what data is available, the output required and how they function. The following section will explain how specific types of machine learning algorithms operate.

2.2.1 Linear Regression

Linear regression represents a good initial starting point for ML algorithms as its simplicity makes it easy to interpret and implement (Montgomery. D. C., 2021). It assumes the existence of some function comprised of a linear combination of the input covariates which map to some continuous output variable.

The basic structure of linear regression for a single case takes the form seen in formula 2.12 where y is the prediction target, x is a row vector of input variables of length d , θ is a row of weightings of length $d+1$ and ε is addition Gaussian noise. The $+1$ represents an additional weighting that is used to calculate the intercept for the samples. When extended to multiple cases, this changes to formula 2.13 where y is now a column vector of prediction targets of length n . X is a matrix of input variables with n sample rows and d variable columns while θ and ε remain the same. The major drawback with linear regression is that actual mapping of the inputs to the outputs may be too complex and result in poor performance, the obvious example being where the underlying function is not linear.

$$y = x\theta + \varepsilon$$
2.12

$$y = X^T\theta + \varepsilon$$
2.13

2.2.2 Logistic Regression

Logistic regression is a classification algorithm that assigns a discrete binary class to each input sample (Kleinbaum & Klein, 2010). Like linear regression, weightings are optimised for each input feature but after this a logistic (logit, sigmoid) function is applied to the result providing a binary split. This scales the initial outputs above 0 to values between (and inclusive of) 0.5 and 1 and outputs lower than zero to between 0 and 0.5. A feature of the sigmoid function is

that as the input values get increasingly close to positive or negative infinity, the outputs get increasing close to either 0 or 1 but never reach either extreme. This allows for a probabilistic binary classification achieved by applying a threshold of 0.5 to achieve a class of either 0 or 1. This can be extended to multiple class output using the 1 vs all method, where a separate model for each class is created and optimised.

The logistic regression equation is shown in formula 2.14, where $S(x)$ is the result of the sigmoid function, e is the exponential of $-x$ and f is the value that is to have the function applied to it. This can then be used to extend the existing linear regression notation. This is plugging into the logistic/sigmoid function in place of f , as seen in formula 2.15 and visualised in Fig 2.4.

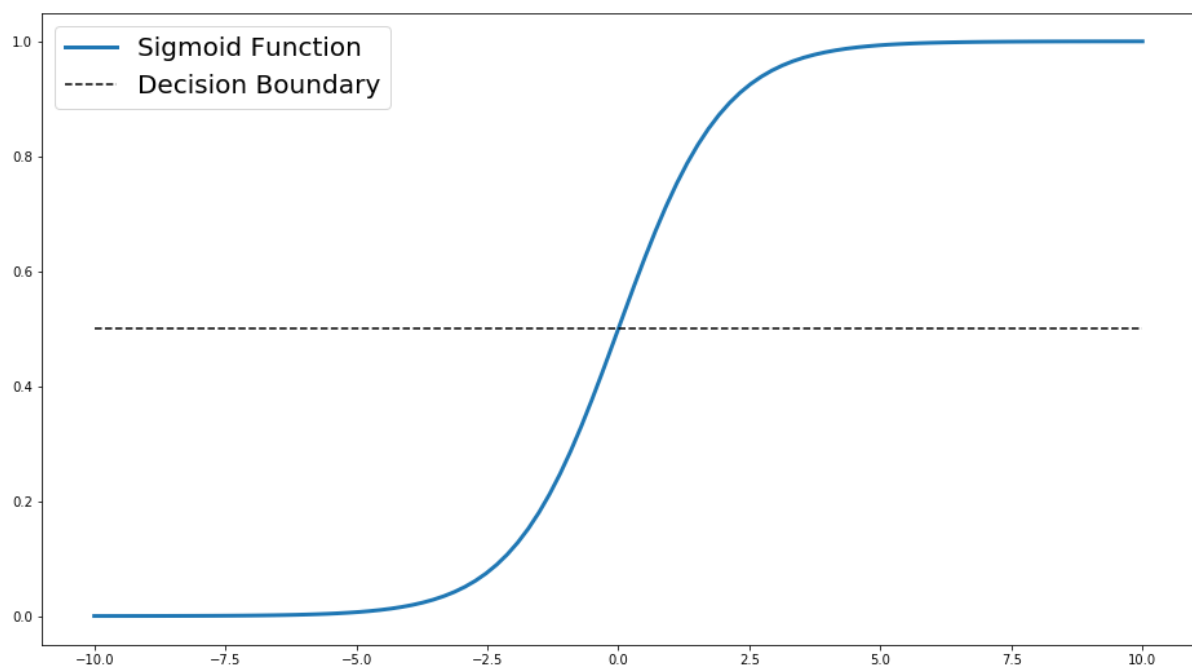


Figure 2.4 Sigmoid Function. The sigmoid function converts a set of real number into a range of values between 0 and 1. Negative values range from 0 to 0.5 with 0 being exactly equal to 0.5 and positive values range from 0.5 to 1. As the absolute values become closer to infinity, they get closer to either 0 (negative) or 1 (positive) but never actually reach these extremes.

$$S(x) = \frac{1}{1 + e^{-f}}$$

2.14

$$y = \frac{1}{1 + e^{-(x^T \theta)}}$$

2.15

Once again, it is a linear interpretation of the data, which may struggle with more complex relationships unless accompanied by additional feature selection. This is offset by its high levels of interpretability which is highly desirable in a medical environment or tool where both clinicians and patients wish to understand precisely how a given conclusion was reached.

2.2.3 K Nearest Neighbour (KNN)

KNN (L. E. Peterson, 2009) attempts to categorise some sample with an unknown output variable, based on the class of the closest K training samples located within some n dimensional input space. For instance, with a K value of one, the target output is assigned the class of the single closest training point. For K values larger than one, the simplest implementation is using a linear combination of the outputs (i.e. the majority).

KNN can be enhanced in several ways. The first is by altering the K hyperparameter as shown in Fig 2.5. The second is the choice of distance measure. Fig 2.5 uses a Euclidean distance measure shown in formula 2.16. Several other measures are possible such as the Manhattan, formula 2.17, and its more generalised form, the Minkowski distance, formula 2.18, where p is a distance scaling parameter.

$$d(x, y) = \sqrt{\sum_{i=1}^n (y_i - x_i)^2}$$

2.16

$$d(x, y) = \left(\sum_{i=1}^n |y_i - x_i| \right)$$

2.167

$$d(x, y) = \left(\sum_{i=1}^n |y_i - x_i| \right)^{1/p}$$

2.8

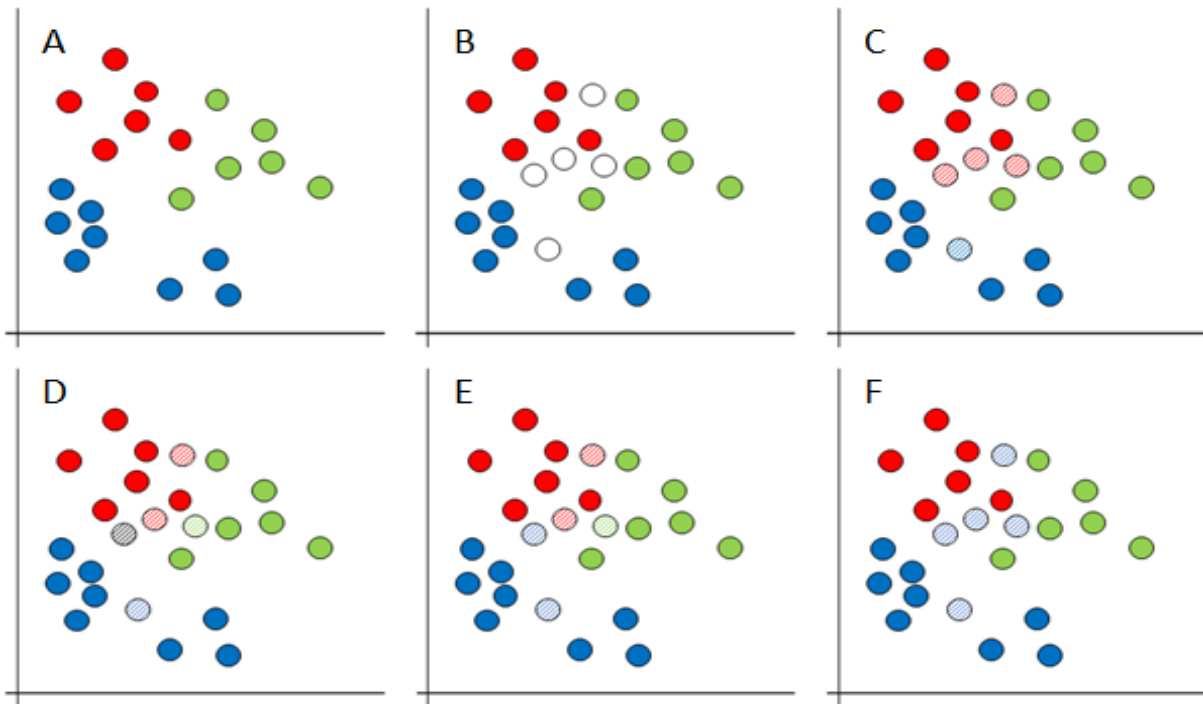


Figure 2.5 K nearest neighbour. A: Training data for 3 classes (red, green and blue), B: 5 new testing points (clear). Classification of samples (coloured diagonal lines) with a k value of 1, 3 and 5 using majority voting are shown in C, D and E. F represents a poor implementation using a K of 20, all points, which will always output blue as there are more blue training points. There is also the final classification. Majority weighting was used previously, but this could be a more complex relationship such as weighting each contributor by its distance from the target.

2.2.4 K-means clustering

The simplest form of K-means clustering attempts to separate samples into different clusters (classes) based on their position within some n-dimensional space (Mannor et al., 2011), Fig 2.6. It first assumes that there are K, specified by hyper-parameter, classes present within the data, although a robust implementation will try several values of K. First, centroids for each class (K points) are placed randomly within the space, then each sample is assigned to a class based on the nearest centroid. The central point of all samples within a given class is found and the centroid for that class moved to this location, followed by reassignment of all samples to their new closest centroid. This process is repeated until either the centroids remain static or after a certain number of iterations has elapsed.

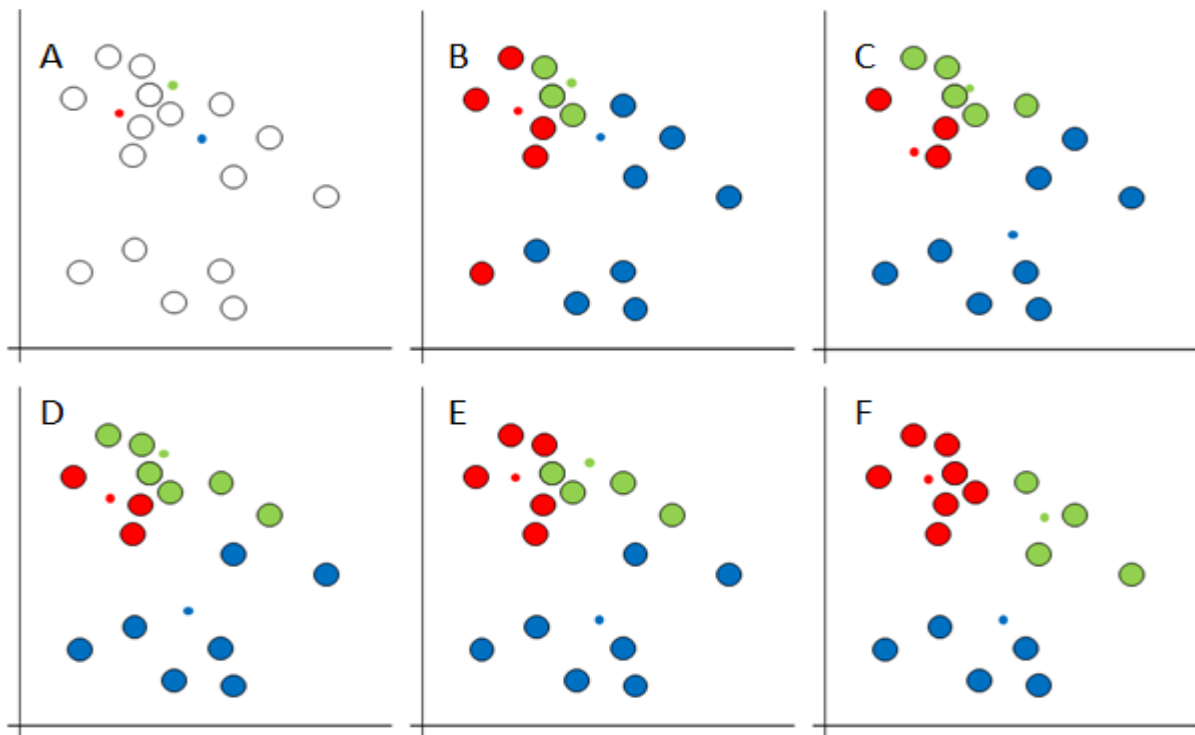


Figure 2.6 K means clustering. A: Raw data without classes and $K=3$ randomly assigned centroids. B: The initial clustering of data where each point is assigned by Euclidian distance to the nearest centroid. C-E: The centroid is relocated to the centre of all samples assigned to its class and the samples are then re-assigned to their nearest centroid. The centroids have reached a position where iteration no longer changes their assigned samples.

2.2.5 Decision Trees and Random Forests

Decision trees (Quinlan, 1986) model output by repeatedly splitting data, Fig 2.7. First a dataset is assigned to an initial root node and these data are split depending on some input variable value (e.g., male or female). These subsets are passed on to branch nodes which perform additional splitting operations that further divide the dataset until terminating at leaf nodes which define the output value or class of a particular sample. A node is considered the parent of any nodes its split directly contributes too, while these nodes are considered the children of that parent. This allows a non-linear relationship between the inputs to be established, there

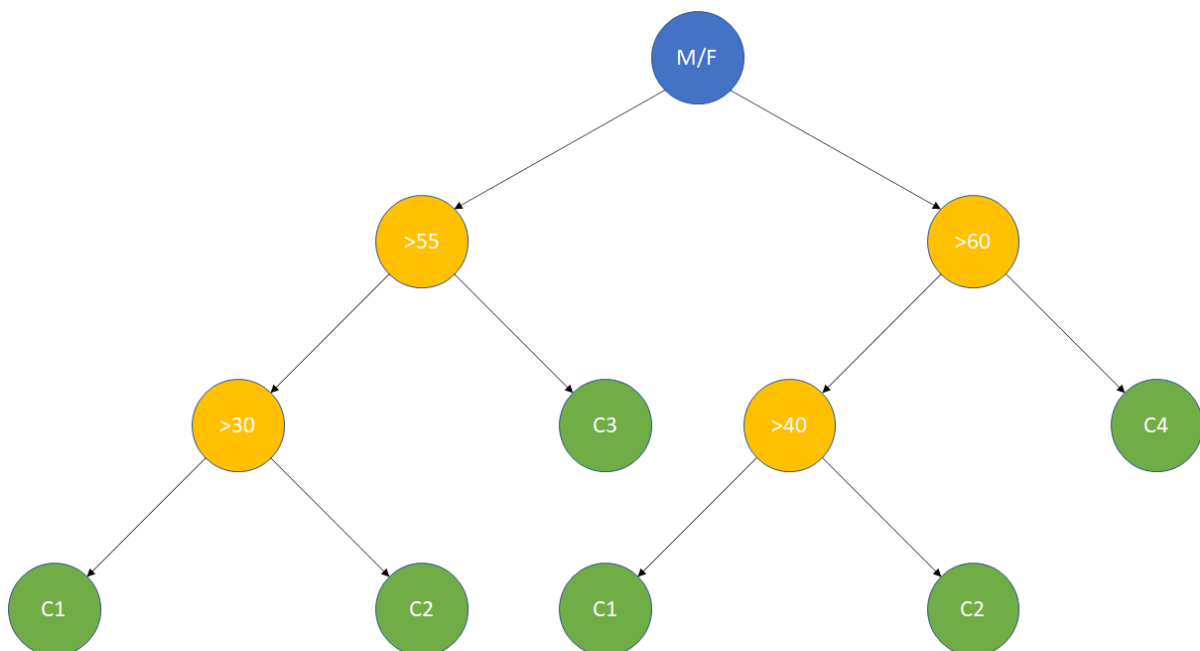


Figure 2.7 Decision tree classifier. The blue root node contains all data and performs the first split, in the example between male and female. The root node is the parent of the first two yellow nodes, the children of the root node. This division process continues, in this instance divisions all by age, until reaching a leaf node (green) which represent the final classifications. Each of the non-leaf nodes is the parent of at least one child node as defined by the arrows in the graph. It is important to note that different leaves can point to the same classes but have arrived via different paths through the tree.

can be multiple paths through the tree which split the same features but at different nodes and with different weights.

A common enhancement to decision trees is their incorporation into a ML method called random forest (Breiman, 2001). This constructs multiple decision trees using a process known as bootstrap aggregation or bagging. Each tree is constructed using randomly selected subsets of the training data, enforcing a limit on how many input features a split can use for separation and allowing each tree to have a different structure. The overall prediction output of a random forest is the average of the outputs from all constructed trees.

Decision trees, and by extension random forests, have the same high interpretability benefits as logistic regression, but can be robust to any non-linearity present within the input features as necessary. They also have the useful ability to show precisely how a decision (classification) has been reached, by viewing the steps taken from root to leaf node within each tree.

2.2.6 Markov Random Fields (MRF)

Markov random fields (MRF) are another widely used ML method and represent relationships as a type of undirected graphical model (S. Z. Li, 2001), Fig 2.8. A model's graph represents the relationship of input variables to each other, where there exist some subsets of variables which are directly linked. These linked variables are grouped together into substructures known as cliques. Each variable is a clique unto itself, as are each pair of linked variables and any group of completely interlinked variables.

For each clique within the graph, a potential function, the probability distribution, is calculated using formula 2.19. Here ϕ_c is the potential function for some clique, y_c is the subset of nodes for that clique and θ_c is the parameters for that clique that need to be optimised. To predict the output for some subject y we sum over the product of all cliques and normalise by a partition

function $Z(\theta)$, formula 2.20. Here the partition function $Z(\theta)$ is the sum of product of all cliques for all possible values of y , formula 2.21.

$$\varphi_c = (y_c | \theta_c)$$

2.19

$$p(y|\theta) = \frac{1}{Z(\theta)} \prod_{c \in C} \varphi_c(y_c | \theta_c)$$

2.20

$$Z(\theta) = \sum_y \prod_{c \in C} \varphi_c(y_c | \theta_c)$$

2.171

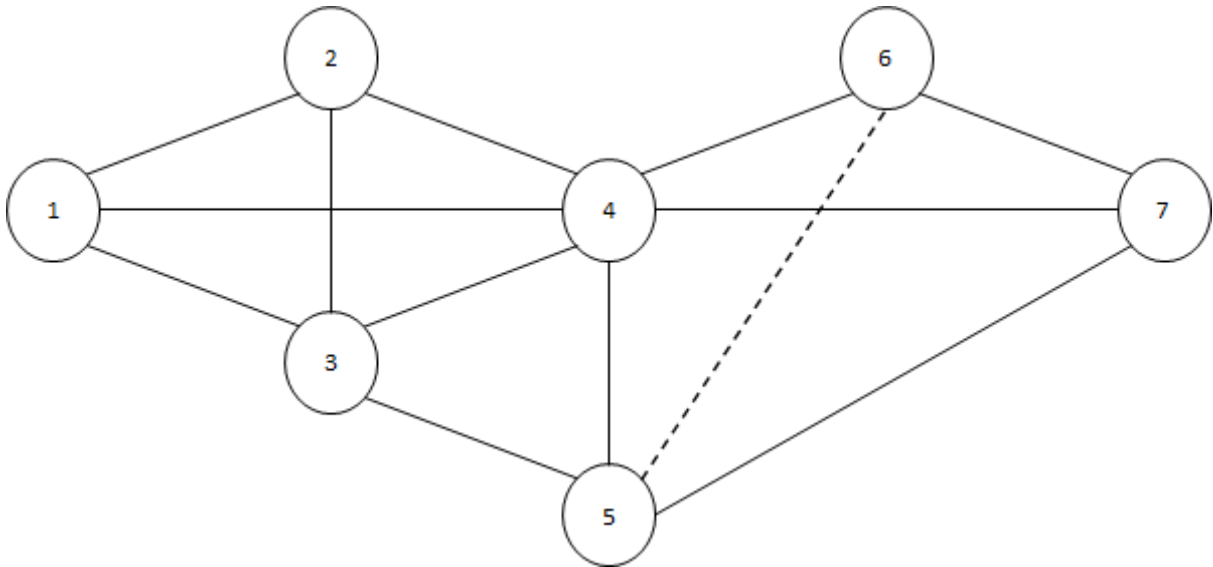


Figure 2.8 Markov random field (undirected graph). Each node $\{1, 2, 3, 4, 5, 6, 7\}$ and pair of linked nodes $\{\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \{3, 5\}, \{4, 5\}, \{4, 6\}, \{4, 7\}, \{5, 7\}, \{6, 7\}\}$ are cliques. Additional cliques include any group of completely connected nodes such as $\{1, 2, 3, 4\}$ (and their sub combinations such as $\{1, 2, 3\}$), $\{3, 4, 5\}$, $\{4, 6, 7\}$ and $\{4, 5, 7\}$. If an additional line is added (dotted line) then $\{4, 5, 6, 7\}$ would now also be a clique.

Intuitively, a 3D image-based graph model could be comprised of all voxels within an MRI image where neighbouring voxels each form pair-wise cliques.

One issue which can arise when using Markov random fields is with the partition function $Z()$ during optimisation and inference. This is because this function needs to sum over of all possible observed variables, for example if there were 100 input variables each with ten possible inputs this would be a 10^{100} summation. Computationally speaking this is expensive and may prohibit its use depending on the number of input variables, such as using all voxel intensities within an image.

2.2.7 Support Vector Machines (SVM)

SVMs are a type of wide margin classifier and while they do serve primarily in this purpose, they can still be implemented as a regression algorithm (Cortes et al., 1995). They operate by trying to apply a binary separation to the data (multi-class can again be achieved using multiple one vs all classification) within the n -dimensional input space. This boundary is fit such that the margin (distance) between one, or more, input points of each class, the support vectors, are maximised. Having many support vectors makes the decision boundary more accurate but comes at the cost of increasing computational expense.

A SVM's decision boundary can be either linear, known as a hard margin shown in Fig 2.9 A, formula 2.22, or non-linear, called a soft margin, Fig 2.9 B, formula 2.23.

$$y_i(w^T x_i - b) \text{ for all } 1 \leq i \leq n \quad 2.22$$

$$\max(0, 1 - y_i(w^T x_i - b)) \quad 2.183$$

Non-linear boundaries are better when the input data has no obvious linear separation between the classes present in the dataset and the samples are not closely intermingled, such as seen in Fig 2.9 B. The boundaries can keep a fixed distance from any support vectors to divide the data but are susceptible to noisy data and more likely to overfit the data, as shown in Fig 2.9 C.

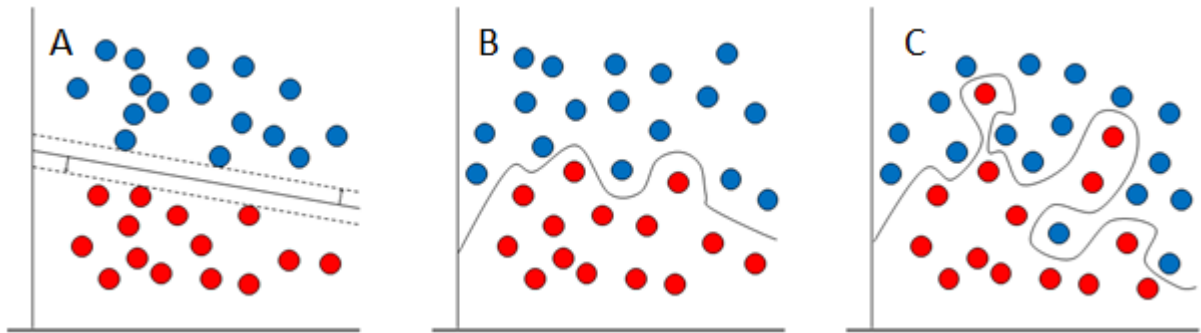


Figure 2.9 Support vector machines. Image A shows a support vector machine (SVM) with a linear kernel. Image B shows an SVM with an RBF kernel, allowing it a non-linear fitting of the data. Image 3 shows an overfitting RBF SVM, where, despite the fact there are obvious outlier in both classes, it has found a decision boundary which still perfectly separates the data, likely resulting in poor generalisability.

It can be enhanced using a method known as the kernel trick. It may be that the input data is not easily separable within its native input space, making classification difficult and prone to errors. A nonlinear transformation can be applied to the data by moving it into some higher dimensional input space where this separation is possible. Applying a hyperplane decision boundary to this new space allows for a much better classification accuracy at a higher computational expensive. There is however a solution, instead of directly mapping the inputs to a higher feature space it can instead map only the dot product of the SVM optimisation algorithm. This has the same effect but is much more efficient.

2.2.8 Neural Networks

Neural network is an all-encompassing term for machine learning methods which create multiple interconnected computational nodes, each of which can be something like a logistic regression function, laid out in an array (Schmidhuber, 2015). The first set of nodes, layer one of the network, use the same initial set of covariates but allow each node to independently optimise its parameters weightings. This allows multiple, one for each node in the first layer, different outputs to be created from the same set of input variables. The outputs from this first

layer can then be used by nodes in a subsequent layer, either another processing layer or the output (i.e., the overall prediction of a single sample) layer. This essentially means each layer is generating new features based on the initial covariates which are more useful for the classification. This process allows for extremely complex functional mappings of input features. An example of a simple neural network can be seen in Fig 2.10.

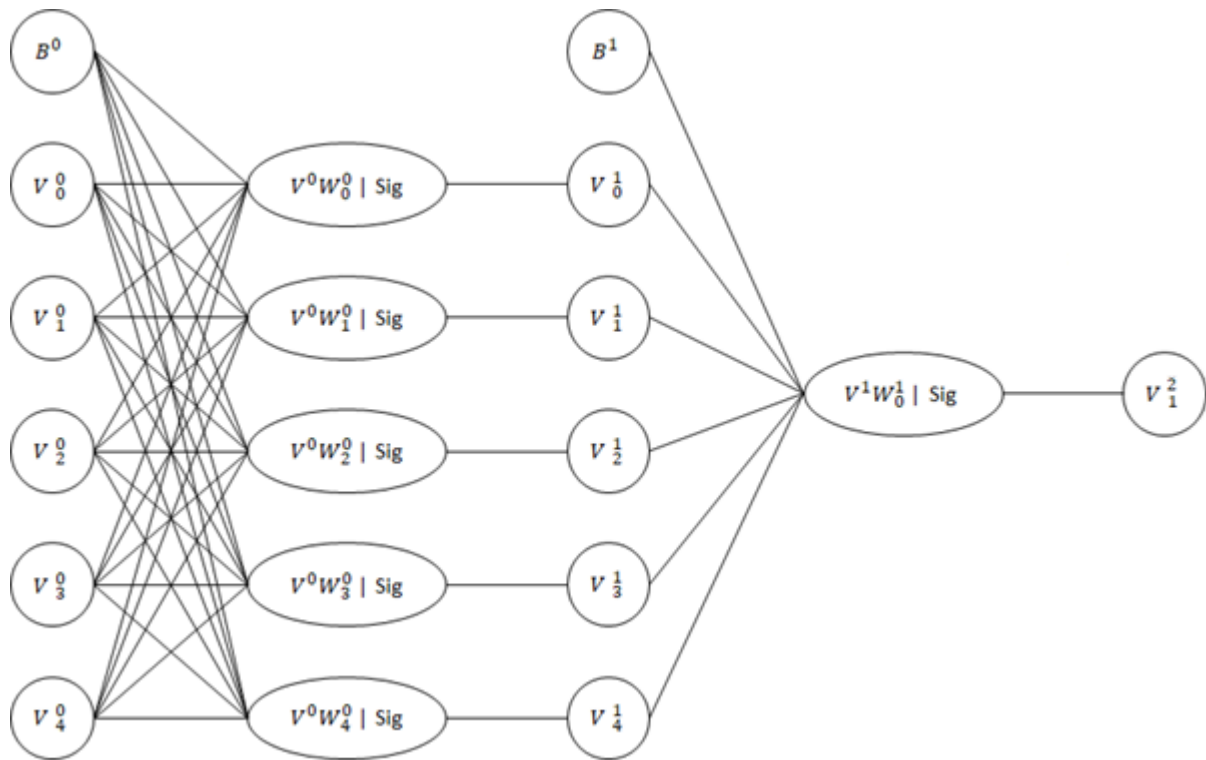


Figure 2.10 Basic neural network architecture. Figure shows a basic implementation of a neural network with relatively simple architecture. Each circle with a V is a variable, where the subscript indicates different variables for a single sample, e.g. height or weight, and the superscript indicates where they are created; the initial variables (0), created by the hidden layer (1) or the final output (2). Each oval is a single processing node that takes in all inputs from the initial variables or previous layers output. Each W is the node's parameter weighting, where the superscript is the layer, and the subscript is the node within that layer. The Sig is the activation function, in this case a sigmoid, that decides the node's final output.

The architecture of neural networks can be highly variable, with differences in the number of layers and nodes, their interconnectability, the activation functions as well as specialised layers.

This results in multiple types of neural networks such as convolutional neural networks (CNN), U-Nets and generator adversarial networks (GANs) to name but a few. The choice of neural networks is highly dependent on the task with CNNs being used widely in imaging processing and GANs seeing use in generation of artificial data.

Neural networks do have limitations. They often require large amount of training data for optimisation, the amount of which increases with the complexity of the challenge. Another limiting factor is the low interpretability associated with neural networks. The complex structure which forms the prediction function can make it difficult to identify exactly how a neural network has reached its conclusion and people may be reluctant to rely on such a tool in a clinical setting. Neural networks also require large times and computational investment. Selection of network architecture is usually specific to a problem and can be highly unique.

Within the domain of neuroscience several types of neural network architecture see common use. CNN's (Krizhevsky et al., 2012) combine convolutional layers to generate feature maps, max-pooling layers to down sample these maps and fully connected layers to generate some output. Autoencoders (Hinton & Zemel, 1993) use contraction layers in a manner similar to a CNN (the encoder) but is followed by expansion up-sampling layers (decoder) which return the output to image space. The frequently used U-net (Ronneberger et al., 2015) is a subtype of auto-encoder designed specifically for segmentation of biomedical images. A GAN (Goodfellow et al., 2020) is used for creation of synthetic images and contains a generator network which produces artificial images, and a discriminator network that attempts to separate these artificial images from real ones. These two networks train and compete simultaneously until the generator produces synthetic images that the discriminator can no longer identify. Recurrent neural networks (RNN) (Medsger et al., 2001) incorporate mechanisms to account for time related points via gates that remember or forget certain features between sections of the network, this makes them useful for studies that include longitudinal data.

2.2.9 Gaussian Processes

Gaussian processes (GPs) make inferences based on the similarity of a target point to a known set of training points and are specified by a mean and covariance function (Rasmussen & Williams, 2006) Fig 2.11. Rather than a single function to describe the mapping of input variables to the output value, GPs instead describe a distribution of functions from which any sampled function could potentially model the data. An initial prior which incorporates any assumptions about the mean and variance of the distribution of functions that fit the data is constructed. The underlying behaviour of the functions within this distribution are described

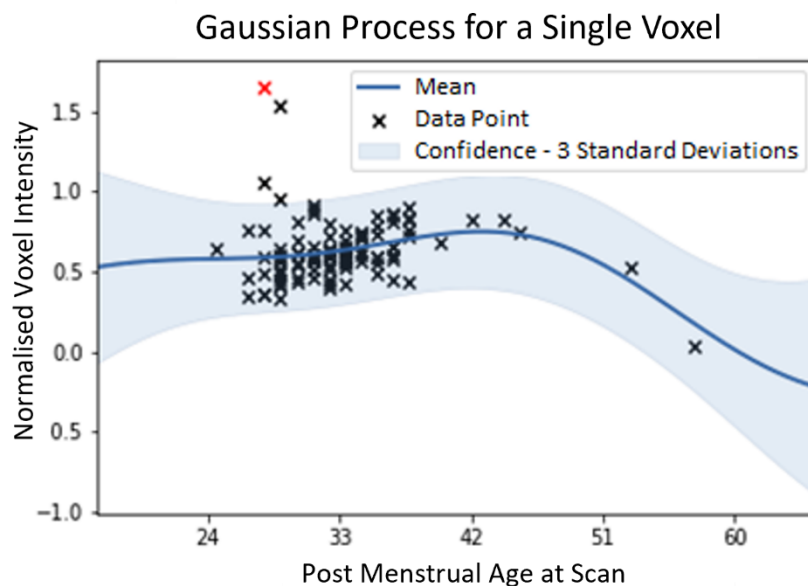


Figure 2.11 Simplified Gaussian process. This figure shows a simple Gaussian process (GP), that takes in age as a variable and plots a single voxel intensity. The black Xs are members of the normative cohort which the GP inferences from. The blue line is the mean of all functions the likelihood has fit to this distribution of points. The blue shaded area represents the 3 standard deviation confidence interval around the mean function, any function that primarily exists within this shaded area could potentially fit the data. Any new sample that fits within this band would be considered normal, a sample outside of this, such as the red X can be considered an outlier.

by a covariance kernel, for example the popular radial basis function (RBF). This assumes that similarity between points is based on the Euclidian distance between them within the n dimensional input space. The GP prior is converted to a posterior by its application to a set of known training points which alter the mean and variance at any given point based on the rules specified by the covariance kernel.

Formally, a GP begins with a prior ε based on a normal distribution N around an initial mean, usually 0, and some variance σ , formula 2.24. Next, some covariance function is chosen for quantification of the similarities between the data and is used to construct a covariance matrix Σ , formula 2.25. In this instance the kernel used for our covariance matrix is the radial basis function (RBF), formula 2.26 where σ is the output variance and l is the lengthscale. Predictions are made based on the joint conditional probability of the training data and a new set of unseen points, formula 2.27.

$$\varepsilon \sim N(0, \sigma_n^2) \quad 2.194$$

$$\Sigma = K(x, x) \quad 2.205$$

$$K(x, x) = \sigma^2 \exp\left(-\frac{d(x, x')}{2l^2}\right) \quad 2.216$$

$$\begin{bmatrix} f \\ f^* \end{bmatrix} \sim N\left(0, \begin{bmatrix} k(x, x) & k(x, x^*) \\ k(x^*, x) & k(x^*, x^*) \end{bmatrix}\right) \quad 2.227$$

Where x is the set of points with known output values, x^* is the set of points with unknown output values, $k(x, x)$ is the covariance matrix training data, $k(x^*, x^*)$ is the covariance matrix

of the testing data and $k(x^*, x)$ and its transposition $k(x, x^*)$ being the covariance matrix between the training and testing data.

An advantage when using GPs is that they work well with smaller datasets. This is because they make their inferences based on similarity their main requirement is that the input space has a sufficient distribution of training samples from which inferences can be made. Additionally, the use of optimisation of the posterior and covariance kernels acts as constraints on the prediction where training samples are sparse. This has the effect of returning the prediction to the mean unlike some other ML methods which attempt to apply a learned function which may not hold for data outside the observed distribution.

2.2.10 Model Parameter and Hyper-Parameter Optimisation

One key task needed by all machine learning algorithms is the selection of their parameters and/or their hyper-parameters. In many instances, such as linear or logistic regression and neural networks, the parameters are the weightings applied to each of their inputs, this can also include linear SVM's optimal plane to separate classes or the length scale of a GP's RBF kernel. Hyperparameter optimisation can include things like the selection of K in KNN or K-means, the architecture of a neural network or kernel selection of a GP.

The selection of these parameters is often achieved by an optimisation algorithm. In this work, parameter optimisation is achieved using the Adam optimiser (Kingma & Ba, 2014), a refinement of stochastic gradient descent ("Encyclopedia of Machine Learning," 2010). This iteratively applies the model to the training data, assesses its performance based on some loss function (for example MSE) then corrects the parameters to minimise this loss function until a stopping criterion is reached. The main advantage of the Adam optimiser is its learning rate hyperparameter, which controls how much a parameter value is changed in each iteration and is traditionally a single fixed value. In the Adam optimiser, a separate learning rate is

maintained for each parameter and can vary as the optimisation progresses allowing it to adapt itself to the data.

2.3 Existing use of Machine Learning within Neuroimaging

The majority of the following papers look at either image-wise abnormality detection or voxel-wise segmentation of tissue types or abnormal areas. Many of the methods reviewed also use combinations of several ML techniques to perform their task.

2.3.1 Lesion Detection: Segmentation of Diseased Tissue

2.3.1.1 Tissue Maps

Srhoj-Egekher et al., 2012 used raw and gaussian smoothed T1w and T2w voxel intensities as inputs to a multi-atlas segmentation method expanded with tissue specific KNN classifiers and corrected voxels with multiple memberships with a naive bayes classifier. Gaussian mixture modelling, fuzzy C-means clustering were combined by Saladi & Amutha Prabha, 2018) to automate tissue segmentations followed by correction using and MRFs. Kumar et al., 2018 combined the benefits of U-net (good recognition of both course and fine features via up- and down-sampling of patch information) and SegNet (computational efficiency and low training needs) for structural segmentation. Delannoy et al., 2020 used a generator network to repeatedly create artificial high-resolution images and tissue segmentations, while a discriminator network attempts to continuously isolate these from the actual high-resolution images and manual tissue segmentation.

The primary issue affecting these methods is that the segmentations rely on voxels having membership to some known class. This allows features associated with that class, such as location or intensity, to be identified and used to assist the process. Atlas based methods, such as Srhoj-Egekher et al., 2012, are not ideal for pathological segmentation as they are highly reliant on location information to make segmentations. Segmentation methods based on

intensity, such as the gaussian mixture models used in Saladi & Amutha Prabha, 2018, are more applicable but suffer from needing prior knowledge of the values associated with given tissue classes.

The architecture of the GAN used in Delannoy et al., 2020 offset a primary disadvantage of neural networks by requiring fewer training samples, however it would still need adaption as due to its requirement of tissue priors. An interesting aspect seen in these methods is that multiple machine learning methods are used to either compliment, correct or compensate for shortcomings of other methods. This includes KNN and naive bayes used by Srhoj-Egekher et al., 2012 and the Fuzzy C-Means and MRF used by Saladi & Amutha Prabha, 2018.

2.3.1.2 Multiple Sclerosis

N. Subbanna et al., 2015 created an iterative, multi-modality model using a voxel-wise Markov Random Field (MRF) incorporating surrounding image intensity and suspected tissue class combined with a region-level MRF where nodes are voxels with the same region and class which accounted for partial volume. Guizard et al., 2015 adapted a type of image denoising that compared patches from a target image to those from healthy control images, a method called non-local mean, applied to multi-modality images. S. Roy et al., 2018 used 2D patches from multi-modalities as input to separate convolutional neural network sections which identify features specific to that modality before combining them in a second convolutional section to estimates overall contribution to the lesion class.

Much like the tissue segmentation methods, the primary issue with N. Subbanna et al., 2015 and S. Roy et al., 2018 is that these methods attempt to detect features specific to MS lesions. This would require either multiple models, one for each potential pathology that is attempting to be detected or a single extremely complex model. In either case it would require a large amount of computational training time and access to large amounts of varied training data.

The use of patches from healthy controls to detect outliers as used by Guizard et al., 2015 is more interesting. Detection of the lesions is based on the difference, specified by the distance of a sample from an expected mean, between a patch from the target and patches extracted from normal controls. As such it is invariant to the type of lesion as well as the direction and magnitude of the intensity difference. This could potentially make it much more generalisable than the previous methods.

One common theme throughout these models is that all of them use multiple modalities to assist their detection. As was noted in the introduction of this thesis, it is common during radiological reviews to view multiple modality types as each contain complementary information allowing for a more complete view of present pathologies and lesion load. As such, automated lesion segmentation that use multiple modalities information are likely to improve on methods that rely on single modality information, especially when attempting to detect different types of pathology.

All three models also utilised patches of voxels during their estimations. In the case of N. Subbanna et al., 2015, combination of voxel-wise MRF based purely on intensity was complemented by the second patch based MRF which introduced structural context. Patches strike a balance between absolute localization, such as seen in purely voxel-wise methods, and incorporation of the surrounding tissue, which at the opposing extreme would be analogous to using the entire 3D image.

2.3.1.3 White matter lesions

Steenwijk et al., 2013 segmented WM lesions in T1 and FLAIR images by incorporating voxel intensity, position and probable tissue class from multi-atlas segmentation tissue priors (Aljabar et al., 2009) into a KNN model. P. K. Roy et al., 2015 created inputs for an initial RF classifier segmentation by using a novel filtering method to enhanced FLAIR image lesion

intensities followed by refinement to the segmentation using MRFs to incorporate spatial information to aid false positives removal. Coupé et al., 2018 isolated the candidate areas-based WM tissue intensity averaged over all subjects using the method suggested by Guizard et al., 2015 which is then corrected using a patch based convolutional neural network to remove false positive.

2.3.1.4 Tumour

Szabó et al., 2018 used a min, max, average, gradient and Gabor values from variously sized multiple modality 2D and 3D patches as inputs to an RFs classifier to perform voxel-wise tumour segmentation followed by neighbourhood weighting correction. Kniep et al., 2019 used radiomic data (first order, shape, texture and their wavelet transforms) as input to a multi-output RF to identify features and segment gastrointestinal, breast cancer metastases, small cell lung, non-small cell lung and melanoma tumour.

The tumour segmentation methods placed greater emphasis on identification of features within an MRI context known to be associated with that pathology. This is understandable as while previous segmentation tasks, healthy tissue, and WM lesions, can be associated with modality specific intensity values, tumours have a more varied appearance in terms of shape, location and intensity values. While interesting in this context, for the more generic abnormality detection attempted by this project this would most likely prove counterproductive as these features are unlikely to be transferable between pathologies.

Kniep et al., 2019 is an interesting example as, aside from the healthy tissue classes, this is one of the few methods that attempts to perform multiple simultaneous segmentations. This presents the opportunity of using a single method to identify multiple pathologies but based on a more generic feature subset. For instance, if combined with a method such as that used by

Guizard et al., 2015, different levels of intensity variation could be associated with either different types of pathology or different levels of severity from a single model.

2.3.1.5 Stroke and Haemorrhage

N. K. Subbanna et al., 2019's second contribution detected sub-acute ischemic stroke lesions in FLAIR images by creating an initial purely Gabor texture-based segmentation which is refined via a voxel-wise MRFs using these same textures and voxel intensities. Nazari-Farsani et al., 2020 segmented acute infarct lesions in ADC images using voxel-wise Crawford-Howell T-test between target subject and a group of healthy controls then applying a threshold to the results. L. Liu et al., 2020 applied u-net convolutional neural network to hybrid images created by fusing T2w and ADC images.

The method used by Nazari-Farsani et al., 2020, is like that seen previously in Guizard et al., 2015 although this time at the voxel rather than patch level. In both instances, they operate by comparing some area of tissue suspected to contain pathology giving an abnormal MRI appearance to a group of healthy controls whose tissue has a typical MRI appearance. Both models fit into the category of what is known as normative modelling. Individuals are compared to a representative population group to discern if their position fits within this typical distribution or is an outlier.

Liu et al., 2020 use of a hybrid image created by fusing ADC and T2 weighted MRI is an interesting development. As noted in the introduction of this thesis and supported by their use in the reviewed methods multimodal information is of great benefit to the segmentation process. By combining these two images during a pre-processing step it maintains the benefits of robust detection, T1w for the lesion core and ADC for lesion extent, while simplifying the model to operate using a single image channel.

Liu et al., 2020 also highlighted one of the downsides to using deep learning, neural network, methods mentioned previously, that they frequently require large amounts of data to be truly effective. The data available for MRI based research is frequently limited with cohort numbers ranging from 10 (Srhoj-Egekher et al., 2012) to ≈ 800 (Makropoulos et al., 2018) participants at the higher end. This is compared to the training sets used in many deep learning methods which often number in the hundreds of thousands or millions of samples. Liu et al., 2020 accounted for this by working with individual slices from the 3D MRI imgs and using data augmentation, flipping, rotating and translating the existing samples to artificially increase the cohort size. Even then they still resulted a much lower number samples then traditionally used for deep learning methods.

2.3.1.6 Lesion Detection via Segmentation Review

An examination of the reviewed methods reveals several common uses of ML techniques that can be broadly divided into one of three tasks; feature highlighting, segmentation and correction. That is, some initial method that highlights feature of interest within an image such as Roy et al., 2015's novel filtering method. This is followed by an initial segmentation method like Steenwijk et al., 2013's multi-atlas segmentation. This segmentation is then refined, for instance using Coupé et al., 2018's convolutional neural network, to remove noise and false positives or improve segmentation coverage.

Examples that are composed of only a segmentation method include Delannoy et al., 2020, Kumar et al., 2018, Guizard et al., 2015, S. Roy et al., 2018, Kniep et al., 2019, Nazari-Farsani et al., 2020 and Liu et al., 2020. Feature highlighting and segmentation was used by Steenwijk et al., 2013 and segmentation followed by correction/refinement was seen in N. Subbanna et al., 2015, Coupé et al., 2018, Szabó et al., 2018 and Subbanna et al., 2019. All three steps can be seen used in Srhoj-Egekher et al., 2012, Saladi & Amutha Prabha, 2018 and Roy et al., 2015.

A popular choice for correction was use of MRF as seen in Saladi & Amutha Prabha, 2018, N. Subbanna et al., 2015, Roy et al., 2015 and Subbanna et al., 2019. The graphical model looks at the classes of the targets voxel's linked nodes, its surrounding voxels, to assess how likely its initial class prediction is. If enough of the surrounding voxels are different, such as if it were a noisy (false positive) prediction, this will be corrected by altering the target voxels class to that of the surrounding nodes.

The methods used by Nazari-Farsani et al., 2020, and Guizard et al., 2015 appear promising as unlike the others they do not rely on features associated with a specific pathology but instead deviation from some expected norm. This could potentially make them highly generalisable to multiple pathologies. The issue is that if these, and indeed any method, were to be applied in a neonatal setting, accommodation will need to be made for the changing appearance of the brain seen during this time of rapidly developing. For the suggested methods, this would require the development of multiple models or limiting the training set to very specific, weekly, time points.

In terms of efficacy, several previous methods showed good performance. Srhoj-Egekher et al., 2012, Kumar et al., 2018, N. Subbanna et al., 2015, Steenwijk et al., 2013, Roy et al., 2015, Coupé et al., 2018, and Subudhi et al., 2020 all showed segmentation sensitivity and dice scores near or over 0.7.

2.3.2 Categorical Disease Classification

2.3.2.1 Multiple Sclerosis

Jog et al., 2015 classified images as either HC or having MS WM lesions by feeding multiple MRI modality patches and positional information into multi-output RF. Zurita et al., 2018 used linear SVMs to classify different types of MS (Relapsing-Remitting, expanded disability status scale (EDSS) > 1.5 and EDSS ≤ 1.5) and HC. They applied Fischer score (He et al., 2005.)

feature selection to fractional anisotropy (FA) values, individual and combined diffusion tensor imaging (DTI) connectivity matrix and rsfMRI correlation matrix.

Saccà et al., 2019 used ICA to identify 15 functional networks, mean signals and MS attributed signal deviations as input to several ML methods (RF, SVM, NB, KNN and NN) where each had their own feature selection. Pinto et al., 2020 modelled MS progression using SVM, decision trees, linear regression and KNN(3), where each had multiple models with increasing amounts of updated medical information to identify relapsing-remitting MS to secondary progressive MS patients and predict disease severity at six and ten years. Schiavi et al., 2022 applied robust feature selection to connectome scores, node strength (summed node link weights), local efficiency (node neighbourhood average inverse shortest path length) and whole net scores, from structural and DTI tractography lesion segmentations as input to KNN, RF, linear-SVM, RBF-SVM and NN for MS detection.

Most of the reviewed segmentation methods focused only on isolating a single pathology or on well-defined tissue classes (frequently GM, WM and CSF). In comparison, even within this subset devoted to MS, many papers attempt multiple types of classification or construct multiple models with different combinations of input features. Examples include Saccà et al., 2019 who used multiple types of classifier and different functional networks as inputs, and Pinto et al., 2020 who created increasingly complex models by appending additional medical information from subsequent longitudinal sessions into the input features.

2.3.2.2 Alzheimer's Disease

Dimitriadis et al., 2018 used an RF to classify brains as HC, MCI, MCI-AD and AD using age, mental state examination score and structural features extracted from T1w images. Ramírez et al., 2018 used ANOVA and partial least squares (Rosipal & Krämer, 2006) feature selection on cortical and subcortical features along with demographic and clinical information as input

as input to two-stage RF for classification of AD, MCI and MCI to MCI. Alickovic & Subasi, 2020 used a binary RF to separate AD from HC using histogram data extracted from structural MRI.

Sørensen & Nielsen, 2018 used a multi-class ensemble SVM to categorise HC, MCI, MCI-AD and AD where each SVM selects different features optimal to a random subset of training samples, like RF, and classification is made by majority voting of ensemble members. Richhariya et al., 2020 used a method called universum SVM-recursive feature elimination (USVM-RFE) which regularises the RFE by the distribution of the data within the input space (the universum). Multiple binary classifications (HC-MCI, HC-AD, MCI-AD) were constructed using three full and feature reduced GM, WM, CSF voxel maps along with 12 binary classifiers based on WM volume, subcortical volume, cortical thickness and all features from the voxel maps. Chagué et al., 2021 built 4 binary SVMs (early onset AD (EOAD) vs depression, late onset AD (LOAD) vs depression, EOAD vs frontotemporal dementia and LOAD vs frontotemporal dementia) using tissue segmentations from T1w images.

Lahmiri & Shmuel, 2019 investigated the effectiveness of using different combinations of cortical thickness, gyrification index (GI), fractal patterns (cortical ribbon, pial surface, GM/WM boundary) and ADAS scores for binary classification of HC and AD using LDA, KNN, SVM and naive bayes classifiers. Rallabandi et al., 2020 compared linear SVM, RBF-SVM, RF, KNN and naive bayes using regional CT measures to classify subjects into HC, early MCI, late MCI and AD categories and isolate input features were useful for separating these classes. Velazquez et al., 2021 classified subjects with MCI who later go on to develop AD by first using an RF to rank the importance of 13 clinical, demographic and brain volume features. Combinations of the more important features were used as inputs to RF, support vector classification (SVC), Logistic regression and an alternative tree model called XGBoost (T. Chen & Guestrin, 2016) for comparison.

Multiple MRFs were used to predict the progression of Alzheimer's disease by Dyrba et al., 2018 by modelling disruptions to different default mode networks (functional MRI) which were shown to be highly correlated with Moore et al., 2019 used a multi-output RF on longitudinal clinical information, which had different and inconsistent timepoints, and demographic data and MRI to separate subjects into HC, MCI and AD.

Once again, the trend is towards both comparisons of multiple types of models and using varied selections of features is seen when applied to Alzheimer's disease. In this instance, SVMs appeared to be more the more common choice being used exclusively in Sørensen & Nielsen, 2018, Richhariya et al., 2020 and Chagué et al., 2021. In Rallabandi et al., 2020 and Velazquez et al., 2021, which looked at multiple types of classifiers, SVM and RF models demonstrated small increase in performance (2-5%) over other tested models such as naive bayes, KNN, logistic regression.

One issue that was noted during the review of Moore et al., 2019 and Velazquez et al., 2021 was the potential effect of averaging the performance of multiple models and having unbalanced classes. While Moore et al., 2019 was able to achieve a high accuracy this was dependent upon correct identification of the healthy control set which overwhelmed the effect of the much lower performance on the more challenging MCI versus AD models. Velazquez et al., 2021 admitted to an unbalanced dataset, however they chose to correct for this by artificially rebalancing these datasets and demonstrated their model's ability to maintain high levels of accuracy under these more robust conditions.

An interesting takeaway is found in Chagué et al., 2021 where they extracted the most important voxel features from their classifier and formed them into segmentation maps. Useful in and of itself, they also found that these maps could be used to enhance two junior and two senior clinical staff's manual segmentations in all except a single senior depression

classification. This not only improves the consistency of the manual segmentations but validates the choice of features used by the models.

2.3.2.3 Tumour

Fetit et al., 2018 used three RBF-SVM models, one per site, to identify different tumour types using histogram and absolute gradient statistics (grey-level co-occurrence matrix (GLCM) (Haralick et al., 1973) and grey-level run-length matrix (GLRLM) (Galloway, 1975)) of ROIs (tumours) as input features, where each model was trained on one site and tested on the remaining two. Vamvakas et al., 2019 first extracted histogram and texture analysis information (GLCM and GLRLM) from K-Means Clustered and radiologist corrected DTI images. SVM recursive feature extraction (SVM-RFE) was applied to these features after application to multiple modalities, rather than just DTI, to separate high- and low-grade gliomas. Kaur et al., 2019 used Bat (Yang, 2010) enhanced fisher criterion for feature selection of first order statistics, GLCM, GLRLM, Laws Texture Features (Tencer et al., 2012), Gabor and Empirical mode decomposition (Colominas et al., 2014) from tumour segmentations created from high contrast FLAIR mapped to the T1 and T1 contrast images to classify hi- and low-grade glioma using a Fuzzy KNN classifier.

Like the tumour segmentation methods, there is an emphasis on identifying MRI based structural features associated with tumorous tissue, with all three methods opting to use both GLCM and GLRLM as covariates. However much like in the segmentation setting, the over reliance on features associated with a specific pathology are most likely going to be counterproductive to the generalizability towards other types of mythology.

2.3.2.4 Stroke, Haemorrhage, Infarct, and other Pathologies

Forkert et al., 2015 tested several SVMs with different combinations of lesion measures and demographic information as input features to predict functional outcome after stroke infarction.

Charidimou et al., 2017 assessed two types of intracerebral haemorrhage and associated injury patterns using normal and multinomial (a multiclass extension) logistic regression. Ordinal logistic regression (Kleinbaum & Klein, 2010), was used by Yuh et al., 2013 for prediction of short-term outcome of patients with mild traumatic brain injury and by Cicuendez et al., 2019 to assess traumatic axonal injury outcome. Sweeney et al., 2021, used combinations of FLAIR, T1w MRI images and demographic information in conjunction with 1-step (regression only) and 2-step (classification followed by regression) RFs to predict MS lesion age.

The models created by Forkert et al., 2015, Yuh et al., 2013 and Cicuendez et al., 2019 represent a more novel use of classification as instead of focusing on identifying a particular pathology they instead try to predict the effect on outcome given the presence of pathology. This is a more challenging task than simply identifying if a pathology present since both the level of outcome disruption and the association between that disruption and the underlying pathology is often more ambiguous.

2.3.2.5 Categorical Disease Classification Review

Classification sees a greater use of clinical information and features extracted from images, such as lesion volume and texture information, compared to segmentation methods which rely more completely on purely image-based information. The usefulness of this characteristic is dependent on the task being attempted. For classification of neonates into either pathological or those with normal appearing brains it may be of limited use as the features, such as location, size and shape, are likely to be highly associated with specific pathologies, of which there are many, and less useful for generic detection. In comparison, these features may be correlated to with outcome, for example larger lesions likely to result in poorer outcomes and damage to the basil ganglia likely to result reduction to motor control or cerebral palsy.

The pattern previously seen in the segmentation methods of combining several ML methods to complement each other and improve performance can also be seen in the classification tasks. For example, in Ramírez et al., 2018 use of a two stage RF where the first stage is a multi-output RF to identify HC, MCI, MCI-AD and AD, and the second stage attempts to alleviate misclassification of MCI by performing a binary classification on members of the HC and MCI classes.

There also appears to be a trend towards comparison of multiple models, of which SVM, RF, KNN and neural networks appear frequently. The use of RF classifiers appears to have become prominent with Jog et al., 2015, Dimitriadis et al., 2018, Ramírez et al., 2018 and Alickovic & Subasi, 2020 using them exclusively. While this may be coincidental, Saccà et al., 2019's RF models had the highest performance and while Pinto et al., 2020 and Schiavi et al., 2022 found SVM's to be the best models, the former used simpler decision trees and the latter's RF performance nearly matched their SVM models.

Several of the methods showed excellent performance including Jog et al., 2015, Zurita et al., 2018, Saccà et al., 2019, Pinto et al., 2020, Schiavi et al., 2022, Lahmiri & Shmuel, 2019, Rallabandi et al., 2020, Richhariya et al., 2020 Velazquez et al., 2021, Chagué et al., 2021, Fetit et al., 2018, Vamvakas et al., 2019, Kaur et al., 2019, Forkert et al., 2015 all showed classification accuracy, sensitivity or AUC of over 0.7, many even reaching scores over 0.9.

2.3.3 Machine Learning Methods Review

Many of the solutions found, for either task, relied on multiple methods and/or in conjunction with an optimised pre-processing method for feature selection. These methods first isolate some feature within the medical notes or image which relates to the given task. This allows a given ML method to optimise only those features relevant to their task rather than the whole image or complete set of input data.

In terms of a specific method to use for either voxel- or image-wise abnormality detection many methods showed potential for a variety of different reasons. This was demonstrated by Rallabandi et al., 2020 when they tested different models for AD classification and found very similar performance. Their SVM-RBF, SVM-LIN and RF models achieved accuracy, sensitivity, and specificity scores of 0.75/0.74/0.73 0.75/0.73/0.74 0.77/0.75/0.73.

While a high-performance method is obviously desirable, within a medical context interpretability is still crucial as it can give medical staff confidence when making decisions. Deep learning based neural networks show high performance but usually require large amounts of data for optimisation which may be an issue with this project as there is a limit in the amount of data available, unless this can be accommodated using a method such as seen in Kumar et al., 2018.

Another issue with many of the methods is that they have been applied only to adults, subjects whose brains appear relatively similar at a variety of different time points. In contrast, the neonatal period is a time of extreme growth and structural change and any method for detection of abnormalities must be able to compensate for this. Failure to do so could result in high rates of false positives and false negatives dependent on age.

These models are also often tuned and limited in precisely what it is they are trying to detect, often focusing on either one, such as MS or Tumour, or a small set of related issues, such as AD, MCI, dementia. For both the segmentation and classification tasks, each model looks for features within an image, potentially from multiple modalities, which are specific to their pathology. Even with closely related tasks, many papers constructed separate models each as seen in Sørensen & Nielsen, 2018, Moore et al., 2019 and Pinto et al., 2020, last of these being a good example as it is performing the same task but at different time points.

When looking at more generic abnormality detection another more general-purpose method is preferable as the alternative would be to have a plethora of models each specific to a different pathology. An idea on how this may be achieved can be seen in the methods used by Nazari-Farsani et al., 2020, and Guizard et al., 2015. By focusing on the differences present within a target image compared to a known set of normal images the detection can be abstracted away from features specific to a pathology allowing for a model that is applicable to a much wider range of abnormalities.

2.3.4 An alternative approach – Normative modelling

This alternative method is the idea of normative modelling. It has seen increasing use in neonates where comparisons to growth charts and developmental trajectories is common practise as well as for feature extraction and outlier detection. Instead of identifying features associated with a given abnormality or pathology, it instead attempts to model normal behaviour over a range of input variables and tries to isolate deviations. Regarding this project this would be attempting to model the normal appearance of the brain's structure at all locations and over all timepoints during the neonatal period. This could then be used to isolate, potentially even segment, areas of the brain which do not match this expected structure.

This method has already seen some use in both adults and more importantly in neonates. GPs were used by Ziegler et al., 2014 for the purpose of detecting local and global GM abnormalities in elderly patients and to identify individuals likely to suffer from MCI and AD. They create a local model, with inputs; age, sex, total GM volume, total WM volume, total CSF and MRI field strength, and a global model, inputs; age, sex and total inter cranial volume, in order to predict the GM volume of each voxel. Binary classification was tested using the 100 testing HC and MCI-AD (92) or AD (128) using combinations of hippocampal Z-scores (voxel-wise prediction – actual GM volume / square root of the variance), both local and global, and brain volume metrics as inputs. While performance was good across all models, the Hippocampal Z-

scores and total GM volume model achieved the highest AUC for HC vs MCI-AD (0.91) and for HC vs MCI (0.94).

O’Muircheartaigh et al., 2020 used a similar method but applied it to a neonatal development dataset to detect PWML. An important note here is that for this voxel-wise method to be successful it needs to account for all the developmental changes that occur during this period just as this project does. The AUC between detected lesioned tissue and a manual segmentation was 0.894 which increased to 0.951 when combined with cluster enhancement.

2.3.5 Conclusions

Each ML method has advantages and disadvantages making them more or less suitable for the expected challenges this project hopes to address. As such a quick review of the methods seen in the journal papers to assess their potential use is necessary before any final decision is made.

Logistic regression could have several potential uses within this project. Although voxel level modelling could binarize (or potentially categorise using either multinomial or ordinal logistic regression) no method like this has been attempted. More likely, it could be used to categorise whole images using image information, patient notes, or some ML generated features such as seen in Charidimou et al., 2017. Its ordinal extension, used by Yuh et al., 2013 and Cicuendez et al., 2019 could also be used if applied to some form of pathology with varying degrees of severity.

MRFs have been used primarily in segmentations. Their graphical nature is frequently used to incorporate spatial data from patches of pixels or voxels when making predictions about tissue type or lesion presence. They demonstrated good performance in normal tissue segmentation in Saladi & Amutha Prabha, 2018, Dice 0.848 and reasonable performance seen in MS segmentations, N. Subbanna et al., 2015 0.69, and general WM lesion detection, Roy et al., 2015 WML Dice 0.76/0.73/0.61. However, later work by Subbanna et al., 2019 showed lower

Dice scores of 0.582 when applied to stroke. MRFs could be used in an alternative capacity, that of correcting the segmentations by removal of noisy false positives.

Much like logistic regression, RF and SVM based methods both show strong promise when applied to the image wide classification problem, as this is where they saw the majority of their use in the reviewed papers. That said, RFs were used by Roy et al., 2015, Szabó et al., 2018 and Szabó et al., 2018 in both WM and Tumour segmentation tasks suggesting a voxel-wise implementation could still be possible. Each paper's focus on a single ML method espouses their own choice as being successful or at least having performance on par with current state of the art methods. The relatively even performance demonstrated in Rallabandi et al., 2020, Saccà et al., 2019, Velazquez et al., 2021 and others indicates that with an unbiased view any of the methods could represent a viable solution.

Neural networks are in an interesting position. They have shown great promise in any area within which they have been applied, especially in image processing and recognition, and their popularity has increase rapidly in recent years. However, they have notable drawbacks, including their large data requirement and black box nature. There is only a limited amount of data available for this project and the lack of interoperability could limit trust during its application. That said it does not rule out their use as a subsidiary modelling technique. If an initial highly interoperative method could be used to segment abnormality or other useful features, neural networks could be used to assist at some secondary level.

GP's major advantage is their high interpretability, as any predictions are based on the similarity of a target to the sample points making it easy to intuit how decisions are made. They are also useful in situations where there is a limited amount of data, as will be the case in this project. The variance output is also a useful feature as it can intuitively be considered an estimation of prediction confidence. Alternativity, Ziegler et al., 2014 and O'Muircheartaigh

et al., 2020 used it to estimate a range of ‘normal’ tissue intensities rather than just a single value.

Ziegler et al., 2014 and O’Muircheartaigh et al., 2020 also predicted normal tissue intensities rather than identifying features associated with a particular pathology. In these cases, detection of the pathologies, MS and PWML respectively, was achieved by finding differences between the actual and predicted tissue intensities of sufficient magnitude and classifying them as abnormal. Potentially, this could make it usable for all pathologies that have an effect on MRI tissue intensity. Of particular interest is the fact that it was used successfully in neonates by O’Muircheartaigh et al., 2020 acting as a solid proof of concept for this project.

2.3.6 Decisions

There are several directions the project could take, highlighted by the review of recent ML based abnormality detection methods. There is a clear split between methods based on either segmentation or image classification, both of which could have a place within this project. The image-wise classification would be ideal for high level detection of pathological neonates, finding those who are likely to have some form of pathology for review by clinical staff. Segmentation is equally important, both to act as a feature selection for the classification and as a way of localising injury to direct clinicians to areas of suspected injury.

For the segmentation, the voxel-wise GP regression used by Ziegler et al., 2014 and O’Muircheartaigh et al., 2020 would seem to be the best solution. Finding areas where a neonatal MRI deviates from a synthesized ‘normal’ brain based on a healthy cohort should be able to detect a broad range of pathologies in a local area. In O’Muircheartaigh et al., 2020 it also demonstrated the ability to accommodate changes seen in MRI tissue intensities over the perinatal period, something that no other reviewed method could. The limited amount of data

available would actually complement GPs as it would prevent the high, and potentially prohibitive, computational expense that working with larger dataset can cause.

The choice of image classification method is more open. Performance has been shown to vary depending on precisely what they are being asked to detect and exactly how they have been implemented. Additionally, it will also be using the GP based segmentations, or some derivation of them, as a novel input source and while some methods have used similar segmentations their exact effect on the classification method will be unknown until tested.

Logistic regression would be simple to implement starting point which has none the less been shown to have reasonable performance. Once this framework has been implemented the choice of classifier could be expanded to include other methods such as SVMs, RFs and GP classifiers. The task itself could also be expanded, for instance judging the efficacy when looking at specific rather than general pathologies, estimating levels of severity or attempting to predict the effect on long-term outcomes.

2.4 Project goals

In summary of the previous sections, the aim of this project is to be able to provide individualised markers of brain tissue abnormalities. This could lead towards treatment targets and guide individually tailored interventions (i.e. precision medicine). Detecting these markers using neuroimaging methods is hampered by the rapid developmental changes in the brain throughout infancy. In practice, this means sensitivity to neuroanatomical changes and heterogeneity that can either mask true abnormality or incorrectly identifying normal developmental process as being abnormal.

This project will develop and apply several machine learning techniques, principally Gaussian process regression, to model typical brain development using large MRI dataset of 500> neonates. The resulting model will be used to maximise detection of abnormal tissue areas for

individual neonates. This will include the creation of personalised atlas's that are the equivalent to a 4D template of the developing brain of all possible points from 28 weeks PMA (Preterm) to 55 (Late Term) weeks PMA. These individual atlas' will be able to incorporate the rapid growth, change in intensity due to myelination and increasing structural complexity due to gyrification that occur during this period.

Statistical/machine learning methods will be used to produce probability maps for whole brain, regional and voxel locations for each template. These can then be used in conjunction with target scans to provide individual analysis of subject brains and assessment of both the presence and level of abnormalities; i.e. variations outside the expected norm for each voxel and/or region.

The analytical images (Z-score maps) generated by the previous methods will then be used to answer a multitude of clinically relevant questions. The first which will be separation of neonates with lesions or structural deformations from those with no MRI positive recordings, equivocal to triage of the target dataset. The second will be the isolation of neonates with a specific complex pathology, hypoxic ischemic encephalopathy (HIE), made-up of several MRI positive pathological features. This will include separation of HIE positive neonates from healthy control neonates. Lastly this project will attempt to predict the likely outcome, poor versus normal, of the neonates with a confirmed HIE diagnosis.

2.6 Hypothesis and Aims

The reviewed literature has demonstrated that the brain undergoes a complex interweaving of developmental processes during the neonatal period. This makes attempting pathology detection in the perinatal setting extremely difficult, a process further exasperated by the large variety of neonatal pathologies and their heterogeneous appearance. It further shows that computer aided detection has seen promising performance in identification of different types

of pathology in adults. However, these are frequently trained and used to identify only a single often homogeneous pathology. A computer guided abnormality detection method capable of identifying multiple types of pathology whilst accounting the changing appearance of the neonatal brain is therefore highly desirable.

Within my thesis I will apply the normative modelling framework to build models using a series of neonatal cohorts with “typical” brain development across MRI contrasts first demonstrated in chapter 3. These models can be used to generate predicted images, allowing generic brain tissue abnormality detection on independent cohorts of neonates with brain lesions. This will be achieved via creation of abnormality maps that highlight areas of deviation between these synthetic images and a neonate actual MRI (prediction error) and is therefore inferred to be representative of abnormal tissue. These abnormality maps can then be utilised as inputs for classifiers. These include attempting to separate neonates with typical appearing brains from those containing any injury in chapter 4. Separating term born healthy controls from neonates with hypoxic ischemic encephalopathy (HIE) and, from neonates with a confirmed HIE diagnosis, separating those who will have either a typical or poor long term outcomes in chapter 5.

Specifically, this thesis aims to:

Aim 1. Generate individualised synthetic images representative of typical tissue in multiple MRI modalities over the neonatal period (chapter 3).

Aim 2. Create abnormality labels that indicate areas likely to contain lesions or abnormal structure in neonatal MRI (chapter 3).

Aim 3. Construct a classifier that separates neonates with typical brain structure and those with lesions (chapter 4).

Aim 4. Construct additional classifiers that identify neonates with HIE and that predict the likelihood of normal vs poor outcome given a hypoxic ischaemic encephalopathy diagnosis (chapter 5).

I hypothesized that:

Hypothesis 1. A voxel-wise Gaussian process model using a normative cohort of neonates with typical appearing brains can be used to create synthetic typical images for neonates with lesions (chapter 3).

Hypothesis 2. The error between a neonates observed MRI and their predicted image can be used to create a map of abnormal areas which can be used to outline pathological tissue (chapter 3).

Hypothesis 3. The incorporation of local structural information can improve performance of these models and better accommodate tissue heterogeneity (chapter 4).

Hypothesis 4. Abnormality maps can be used to classify neonates with normal appearing brains and those with lesions (triage) (chapter 4).

Hypothesis 5. Abnormality maps can also be used to identify neonates with a specific pathological condition, hypoxic ischaemic encephalopathy (diagnosis) (chapter 5).

Hypothesis 6. Abnormality maps can be further used to separate HIE neonates that are likely to have poor functional outcomes from those that are likely to have typical outcome (prognosis) (chapter 5).

Chapter 3. Automatic Detection of Neonatal Brain Injury on MRI

Identification of neonatal brain abnormalities on MRI requires expert knowledge of both brain development and pathologies particular to this age group. To aid this process, an automated technique to highlight abnormal brain tissue while accommodating normal developmental changes is proposed. A developmental model was trained using 185 T2w neuroimaging datasets from healthy controls and preterm infants without obvious lesions on MRI age range = $27 + 5$ to 51 (median = $40 + 5$) weeks + days post menstrual age (PMA). This model was then tested on 39 preterm subjects with known pathology and manual segmentations, age range = $25 + 1$ to $37 + 5$ (median = 33) weeks + days PMA. Voxel-wise Gaussian processes (GP) were used to model gestational age at birth (GA), PMA and sex against voxel intensity, where each GP outputs a predicted mean intensity and variance for that location. All GP outputs were combined to synthesize a ‘normal’ T2w image and corresponding variance map for a given age and sex. A Z-score map was created by calculating the difference between the neonate’s actual and their synthesized images and scaling by the standard deviation map (SD) derived from the GP variance outputs. With a threshold of 3 SD the model highlighted pathologies including germinal matrix, intraventricular and cerebellar haemorrhage, cystic periventricular leukomalacia and punctate white matter lesions. Statistical analysis of the abnormality detection produces an average AUC value of 0.956 and 0.943 against two raters’ manual segmentations. The proposed method is effective at highlighting different abnormalities across the whole brain in the perinatal period while limiting false positives.

3.1 Introduction

Brain injury can critically affect a neonate’s neurodevelopmental trajectory (Anand & Scalzo, 2000). Detection of abnormalities is made challenging by the rapid growth, myelination and gyrification occurring during this timeframe. Early and accurate detection while

accommodating these developmental patterns is vital to guide clinical interventions, reduce their disruptive effects and potentially improve outcomes. Magnetic resonance imaging (MRI) has seen increasing use for clinical assessment in cases of suspected neonatal brain injury (Chaturvedi et al., 2018; Imai et al., 2018; Kline-Fath et al., 2018). Compared to the more ubiquitous Ultrasound, MRI can produce high resolution 3D images of the neonatal brain with excellent signal to noise ratio. These highly detailed images allow for identification of features associated with normal variation as well as abnormal brain tissue and pathologies, even in more subtle injuries, thereby facilitating accurate diagnosis and treatment.

As was highlighted in the introduction of this thesis, examination and review of these images requires expert knowledge of both normal neonatal brain appearance and the presentation of potential abnormalities. This is complicated by the fact that brain regions can exhibit differences in both volume, shape and exact position between neonates. Further, neonatal developmental processes cause these to change over time and at a rate which can vary between individuals. Potential brain injuries are also highly heterogeneous in shape, size and position, even within a single type of pathology. They also have a different visual appearance depending on the choice of MRI contrast such as T1w, T2w and diffusion weighted images.

The result is that even radiologists with extensive training and experience demonstrate both inter and intra-rater inconsistencies when interpreting MRI findings (Bogner, 2018). These reviews can also be time-consuming and expensive, especially neonatal reviews which require staff with additional training and experience to assess them. Some method of highlighting areas of suspected abnormality could assist and streamline this process, reducing staff and hospitals burdens and potentially improving identification consistency.

Computer aided detection methods have been explored in adults for pathology such as stroke using a random forest classifier (Subudhi et al., 2020), Alzheimer's disease (Richhariya et al.,

2020) and multiple sclerosis 2021 (Merzoug et al., 1 C.E.) both using support vector machines. Unfortunately, there is a clear scarcity of similar detection methods for use in neonatal neuroimaging and none that are capable of detecting generic abnormality. Those that do exist are based on segmentation methods transferred directly from use on adult MRI. Ghosh et al., 2014 compared the efficacy of three segmentation methods previously used in adults to detect ischemic arterial stroke and only published the results of their animal models not their neonatal cohort application. L. He & Parikh, 2013 used Markov random field spatial fuzzy segmentation based on T2w relaxometry to detect diffuse excessive high signal intensities (DEHSI). Their later work (H. Li et al., 2019) was also applied to DEHSI but utilised deep learning instead.

A first broad approach to abnormality detection is based on identifying known features associated with a specific pathology within a specific MRI contrast. An example is seen in He & Parikh, 2013 which segmented DEHSI based on the disparity of lesion intensity values within the image. Pathologies such as multiple sclerosis lesions, haemorrhage or tumours also have abnormal (hypo-/hyper) intensity values, positions, shapes or other features which can be learned in order to assist in their identification.

Some studies have only single pass applications, such as Ghosh et al., 2014 who tested three separate methods, symmetry-integrated region growing, hierarchical region splitting and modified watershed segmentation, to segment detected hypoxic ischemic injury. Others, have separate steps such as lesion highlighting (Vamvakas et al., 2019) and noise reduction (N. K. Subbanna et al., 2019) which may utilize a different machine learning technique, K-means clustering and Markov random fields respectively. These are then either directly combined or used the output from one method as input to a second. Other tasks include feature selection, Bahadure et al., 2017 who used Berkley wavelet transform to assist tumour detection), regularizing Pereira et al., 2016 using conditional random fields to assist general tissue

segmentation) and image classification Ramírez et al., 2018 using random forest to identify Alzheimer's disease).

These methods are typically applied only to a single or a limited selection of pathologies or purposes (tumour in Bahadure et al., 2017 and tissue segmentation in Pereira et al., 2016) which would require multiple models to detect multiple types of pathology or tissue. A neonate with both a cerebral haemorrhage and periventricular leukomalacia, two pathologies with very different MRI appearance, would require two separate models each optimized to a specific pathology. Pathology specific detection solutions are often compounded by a requirement of large quantities of labelled training data for each pathology (Rezaei et al., 2017).

An alternative general approach to be agnostic to pathology, is the normative modelling perspective which attempts to characterize what is typical for a given setting, e.g., image contrast for a given population, and therefore base detection on deviations from these expected values. More precisely, locating intensity differences of sufficient magnitude such that they are considered abnormal for a given age and contrast. Within an MRI context, this would propose creation of synthetic images which are representative of “typical” neonatal brain structure based on those observed in a normal appearing group. This allows isolation of any areas within a new neonate's observed MRI that deviates significantly from these predicted images.

Methods that use this synthetic image technique include Bowles et al., 2017 who used voxel-wise kernel regression and X. Chen & Konukoglu, 2018 who used a generator adversarial network. They then calculated the difference between actual and synthesized ‘healthy’ MRI images to detect small vessel disease and tumour respectively. In a similar vein, Azami et al., 2016 calculated Z-score differences between predicted and actual extension and junction maps (derived image features taken from a T1w image) for epilepsy classification. Chen et al., 2018 created their segmentations by applying a threshold directly to these raw differences. Azami et

al., 2016 and Bowles et al., 2017 refined these initial methods using gaussian mixture models and support vector machines, to assist their detection ability.

Recent work from our group by O’Muircheartaigh et al., 2020 used a similar technique based on Gaussian processes (GP), first developed in imaging by Ziegler et al., 2014. In both cases, they constructed multiple, one per voxel, models that predict a ‘normal’ range of expected intensity values at each voxel based on some training samples. The outputs of these models were reconstructed into synthesized images and compared to each subject’s actual brain MRI. Areas within the subjects’ actual images that had intensity values outside of an expected range of values were highlighted to detect Alzheimer’s (Ziegler et al) or neonatal punctate white matter lesions (O’Muircheartaigh et al).

The first task of this thesis is to develop this method for a neonatal brain injury dataset, but one which contains multiple forms of pathology. This will act as a proof of concept for this voxel-wise GP method to be used as a general, rather than pathology specific, abnormality/injury detection over the perinatal period. There will be an emphasis on preterm neonates as the changes in tissue intensity are more dramatic over this extended time period (20 weeks), testing the model’s robustness and ability to accommodate these changes over time. This ability to differentiate between changes seen in typical development and representative of injured tissue is crucial to any method applied to this time frame.

The aim of this chapter is to generate individual maps of abnormality for a set of preterm neonates with known, reported and heterogeneous MRI positive lesions based on a derivation of the method first present in Ziegler et al., 2014. These abnormality maps are then used to create segmentations of clinical lesions and tissue with abnormal intensities found within each individual neonates MRI image. These segmentations are then compared to manual segmentations in order assess the efficacy of the proposed method.

3.2 Methods

3.2.1 Sample and dataset

Ethical permission for this study was granted by the Hammersmith and Queen Charlotte's and Chelsea Hospital (QCCH) Research Ethics Committee. Written parental consent was obtained for each infant. The MRI dataset used in this thesis chapter is composed of 423 preterm neonates acquired at Hammersmith Hospital on a Philips 3T scanner. Ethical approval was granted by a research ethics committee and written parental consent was obtained in each case.

In this analysis, only T2-weighted (T2w) images were used, reflecting the most common imaging type collected in neonatal imaging. Moreover, T2w, and the closely related FLAIR images (T2w images with water suppression), have been highlighted as being particularly useful for review of neonatal brain images (Counsell et al., 2003.; Rutherford et al., 2010).

Quality Control: A comprehensive manual review of all images in conjunction with available demographic information was undertaken by a professor of perinatal imaging and myself. This assessed image quality by locating those images with inconsistent slice intensities, strong inhomogeneity artifacts or having motion blur. This process was also used to identify neonates whose MRI contained some form of lesion or tissue abnormality that could be seen on a T2w image. Of the 423 images, 48 were discarded due to high levels of motion, incomplete demographic data or having some kind of MRI artifact.

The images within the dataset were acquired as either standard clinical T2w or Dynamic T2w images, an image type with T2w appearance but with a much faster acquisition time that is less susceptible to motion, Table 3.1. While individual tissue types may have differences between the two acquisitions, the relative intensities between the tissue types in each acquisition are close enough to be considered interchangeable.

Table 3.1 Acquisition parameters.

Scan	TR	TE	Voxel Size	Subjects
T2 Fast Spin Echo	7000<8000ms	160ms	0.859 x 0.859 x 1mm	347
T2 Dynamic	20000<30000ms	160ms	0.859 x 0.859 x 1mm	28

Of the 375 good quality images, 185 were found to have little or no lesioned tissue on T2w contrast images. Of those images that had small lesions, lesions were almost indistinguishable on the T2w contrast and heterogeneous in location. These 185 images were used as the normative modelling cohort to both optimise the GP model hyperparameters and as inference points during generation of the synthesized normal images.

The remaining 190 images with clear tissue pathology were used as the test set. This included a subset of 39 preterm neonates with an accompanying ground truth segmentation. These were selected based on the range of injuries present in the MRI. This included intensity differences (both hyper and hypointense compared to typical tissue), different locations (subcortical WM, ventricles, cerebellum), sizes, and shapes. Two raters, who were allowed to discuss their findings but kept their delineations independent, manually outlined tissue injury labels using FSL eyes (McCarthy, P. 2023). A visual inspection was made of the remaining 151 images without ground truth segmentations to assess the lesion coverage.

This held-out set of 39 subjects with manually labelled data included the following pathologies seen in neonates: Germinal matrix haemorrhage (N=14), inter-ventricular haemorrhage (N=7), cerebellar haemorrhage (N=11), haemorrhagic parenchymal infarct (N=4), temporal horn cysts (N=1), sub-arachnoid cysts (N=2), pseudocysts (N=1), cystic periventricular leukomalacia (N=3), large(>8 voxels)/many(>3) punctate white matter lesions (PWML) (N=8) and small(<=8 voxels)/few(<=3) PWML (N=10). The greater number of pathologies compared to

individual subjects represent the fact that a single neonate may have multiple MRI positive pathologies present within their image. Each of these pathologies is described in the introduction, alongside their appearances in different MRI acquisition modalities. Subject information is displayed in Table 3.2.

Table 3.2 Cohort information.

Parameter	Value
Total Subjects (Female)	375 (179)
Post Menstrual Age at Scan (PMA) Range (Weeks+Days) [Median]	25+1 to 55+0 [37+1]
Gestational Age at Birth (GA) Range [Median]	23+2 to 42+0 [29+5]
Healthy Controls + Preterm without Lesions - Training	185
PMA Range [Median]	27+5 to 51+0 [40+5]
GA Range [Median]	23+2 to 42+0 [30]
Preterm with Lesions - Testing	151
PMA Range [Median]	25+2 to 55+5 [34+2]
GA Range [Median]	23+3 to 38+6 [29]
Preterm with Lesions and Ground Truth – Statistical Testing	39
PMA Range [Median]	25+1 to 37+5 [32+4]
GA Range [Median]	24+3 to 35+2 [30+3]

There is a deviation between the PMA distribution of the training and testing sets due to a high number of preterm neonates with normal appearing brains imaged at term age. There is still 37

images that cover the period between 27+5 to 37 weeks PMA from which intensity values can be inferred. Although not ideal, this is less of an issue due to the specific model implementation used, which will be covered in section 3.2.3.

3.2.2 Pre-processing

The pre-processing pipeline consisted of 4 steps. Each MRI image first underwent B0 inhomogeneity correction using ANTs N4BiasFieldCorrection (Avants et al., 2014) and removal of extra cortical tissue using FSL brain extraction technique (BET) (Smith, 2002). This was followed by registration between each neonatal MRI image and a 36-week-old template, as this central age between the minimum and maximum age at scan, created as part of the developing human connectome project (dHCP) (Makropoulos et al., 2018). This registration process was completed in two-steps, first an affine registration using FSL FLIRT (Jenkinson et al., 2002; Jenkinson & Smith, 2001) then a nonlinear registration using ANTs Registration (Avants et al., 2014). No image smoothing was performed as it could affect the ability of the model to detect smaller, more subtle lesions.

Another frequent preprocessing step when utilising machine learning algorithms involves normalisation or standardisation of the data being operated on. The choice of which of these two methods would be most appropriate for this task was at the time unknown. As such, both methods were implemented allowing for their comparison. Normalisation involves the subtraction of the data mean from all values and then scaling by the range of those values, formula 3.1. Standardisation also involves subtraction of the data mean but scales by the data standard deviation instead, formula 3.2. As the mean can sometimes be biased by extreme values it is often replaced by the median, allowing for a more robust estimation above average. As such, two additional implementations based on subtraction of the median instead of the mean were also investigate.

$$Normalisation = \frac{X - \mu}{\max X - \min X}$$

3.1

$$Standardisation = \frac{X - \mu}{\sigma}$$

3.2

3.2.3 Modelling Voxel Intensity Values using Gaussian Processes

A detailed overview of GPs and their operation can be seen in chapter 2 of this thesis. For this project, a voxel-wise GP implementation, developed using the GPyTorch library (Gardner et al., 2018), was used to model intensity values in T2w MRI images. The design matrix, input samples and covariates, were composed of the gestational age at birth (GA), the post-menstrual age at scan (PMA) and sex of a neonate, formula 3.3. These were identical for each voxel GP model. The difference between the models is the output variable (voxel) which is composed of the intensity values found at the voxel's location for each neonate. This allows each GP to tailor modelling only those values specific to a given location within the neonatal brain.

$$Sample(x, y) = ((ga, pma, sex), intensity)$$

3.3

For the actual modelling, estimation of posterior hyperparameters, only the cohort with normal appearing MRI were used. This limits the intensity values each GP can inference from. Specifically, only those seen in non-lesioned tissue. A prior ε (an initial estimation of the intensity distribution of voxel n) is specified based on a normal distribution N with a mean (usually 0) and variance σ^2 to constrain the distribution, formula 3.4.

$$\varepsilon \sim N(0, \sigma_n^2)$$

3.4

A standard RBF kernel, formula 3.5, was selected as the covariance matrix, formula 3.6, combined with a Gaussian likelihood. Here, $d(x, x')$ is the Euclidian distance in n-dimensional space between point x and point x' , l is the lengthscale which defines how close within that space points must be to be considered similar and σ is a scaling hyperparameter that controls the amplitude of the functions that fit the data. These hyperparameters were optimized using the Adam optimizer (Kingma & Ba, 2014) covered in chapter 2 section 2.10.

$$K(x, x') = \sigma^2 \exp\left(-\frac{d(x, x')}{2l^2}\right) \quad 3.5$$

$$\Sigma = K(x, x') \quad 3.6$$

One of the primary benefits of the RBF kernel is that it can accommodate the uneven distribution of PMA observed between the training and testing datasets. The kernel looks at only those samples that are closest to the prediction target within the input space when making an prediction. Even though the distribution of samples ages is different between the sets, if there is a sufficient number of samples at any given age an inference is still possible.

Predictions were made based on the joint conditional probability of the training and testing samples, formula 3.7.

$$\begin{bmatrix} f \\ f^* \end{bmatrix} \sim N\left(0, \begin{bmatrix} k(x, x) & k(x, x^*) \\ k(x^*, x) & k(x^*, x^*) \end{bmatrix}\right) \quad 3.7$$

It is common practice in GP kernel selection to accommodate variance in the output values using an additional noise kernel. This is not necessary in this particular implementation, as the

GPyTorch library includes a gaussian noise function as an integral part of the Gaussian likelihood.

An important factor to note is that the noise captured by the model is crucial to successful estimation of voxel intensity which can be attributed to three factors. The first is that, even after standardisation or normalisation, there can be small differences in tissue intensity values between individual neonates which is made up of both true individual differences and image acquisition noise. Secondly, even with good registration performance, there can be some variation in tissue boundary location representing true anatomical variability. Finally, at tissue boundaries there is a partial volume effect where multiple tissue types are contained within a single voxel, making the final intensity value a combination of those present. Any of these situations can result in a GP having output values that can be significantly different between two individual samples. For example, a voxel on the cortical surface boundary for one neonate may contain GM giving a low intensity value while a second neonate may have a voxel which contains high intensity CSF, dramatically different values on a T2w image, as seen in Fig 3.1.

As part of model optimisation, the model adjusts its mean (function) output to a value which minimizes the error between predicted and the actual intensity of the training samples. Alongside this, the variance expands to encapsulate possible training intensities observed for that voxel, assuming a Gaussian distribution. A simplified example of this is shown in Fig 3.1, which shows the variance difference between a GP in the deep grey matter, and one located on the cortical surface. Fig 3.1b shows a wider confidence region around the mean as the intensity values in the variable cortical surface can be quite different, high intensity CSF and low intensity tissue values. Meanwhile Fig 3.1c has a much tighter confidence as it will observe relatively similar deep grey matter voxel intensities.

These single voxel examples can be built upon to demonstrate how GPs can be used to model an entire neonatal brain. Models of single voxels can be combined into entire image volumes. Each of these GPs take in the same GA, PMA and Sex and their mean output is the given voxel's intensity, inferred from the training set, for those covariate values. These predictions essentially synthesize a single typical neonatal MRI image. Using a similar process, a variance image can also be constructed from the GP variance output.

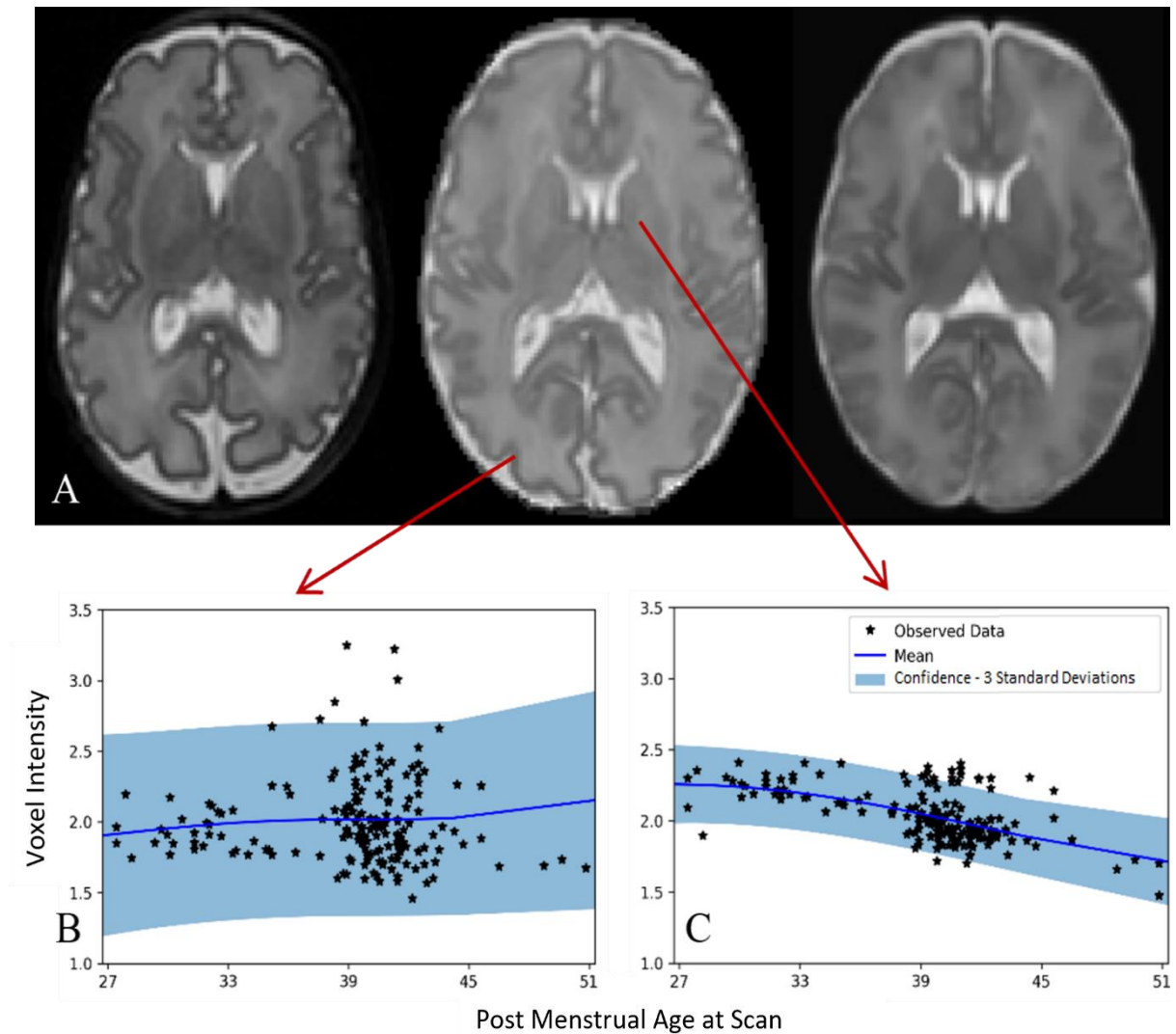


Figure 3.1 Preprocessing and GP performance. A: Raw subject image, subject image after registration to template space and the template itself. B: A single, high variability voxel's Gaussian Process for all training subjects. C: A single, low variability voxel's Gaussian Process for all training subjects.

3.2.4 Detecting Pathological Tissue as Outliers of Typical Development

Estimation of generated lesion label performance compared to manually outlined lesions (ground truth) was quantified using the following criteria: sensitivity (recall), specificity, precision, receiver operating characteristic (ROC), precision recall curve (PRC) and area under curve (AUC) of the ROC and PRC. In order to assess how accurate the synthesized typical images were compared to the actual MRI images, additional testing was completed on the non-lesioned preterm images. These underwent 5-fold cross-validation to construct synthesised images and the difference between these predictions and the raw images were measured using mean absolute error (MAE) and false positive rate.

It is possible to perform abnormality detection based on the absolute difference between a neonate's actual and synthesized image as demonstrated by Bowles et al., 2017 and Chen et al., 2018. The issue is that the difference between normal and abnormal intensity will vary between different voxels and would need to be optimized for each location. Again, while this is possible using a second voxel-wise machine learning operation, an alternative solution exists thanks to the variance images created in tandem with the synthesized images. By first converting the variance to standard deviation image it is possible to construct a voxel-wise Z-score map of the neonate's brain via the standard Z-score calculation, formula 3.8.

$$Z = \frac{x - \mu}{\sigma}$$

3.8

Where x is the actual voxel intensity, μ is the predicted typical voxel intensity output by the GP and σ is the standard deviation converted from the corresponding GP variance value output. The benefit here is that a single threshold applied equally over all voxels in the Z-score map should account for the same level of deviation rather than some arbitrary intensity value difference specific to a given voxel.

It is then possible to assign abnormality based on any given voxel having a sufficiently high or low Z-score value. These can then be highlighted within the Z-score image, and on the neonate's actual image via positional correspondence. In practice, highlighting is completed by first converting the Z-score map to absolute values and highlighting all voxels over a given threshold. These are then colour scaled for review.

To act as an uninformed baseline, simple voxel-wise Z-scores were calculated using the same training data to compare model performance. This should be less specific than those output by the GP model which should be influenced by the model input covariates, the mean changing over time. The resulting AUCs underwent pairwise comparisons using non-parametric Wilcoxon tests.

3.3 Results

After training, the model demonstrated accurate prediction of a typical image based on inference from the voxel intensities in the training data. Fig 3.2A shows a mean absolute error map (MAE) of the differences between the observed and predicted control images where high intensity represents areas of high cortical variability and poorer model performance. Areas with the highest MAE scores were most prevalent on the cortical surface, while the subcortical WM and deep GM had lower values. All values in the MAE image were below 1 which, as these images had undergone standardisation, indicates all predicted voxels were less than 1 deviation from their actual values. From a qualitative perspective, there was smoothing around the cortical surface due to its highly variable structure, Fig 3.1. Less variable areas, such as the subcortical white matter and deep grey matter, were sharper and had better definition.

Examples of abnormality detection based on the Z-score map is shown in Fig 3.2B2-5. In these examples, an (arbitrary) threshold of $Z > 3$ was applied to an absolute Z-score map. This was overlaid over the neonates actual MRI image to highlight abnormalities in four different

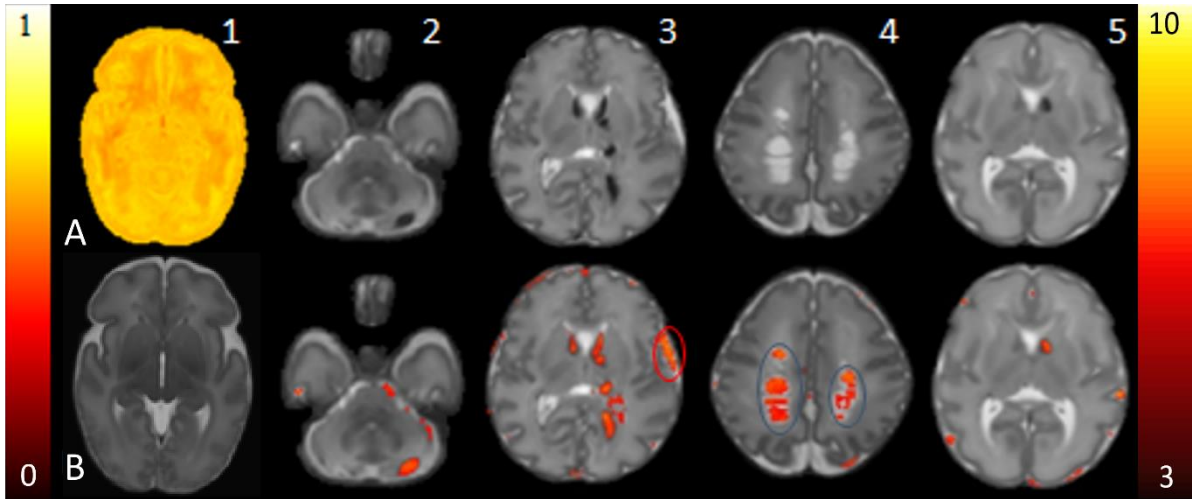


Figure 3.2 Abnormality detection performance. Column 1 row A, mean absolute error map, row B, GP synthesized image. Columns 2-5 row A, pathological images in template space: cerebellar hemorrhage (2), intraventricular hemorrhage (3), cystic periventricular leukomalacia (4), germinal matrix hemorrhage (4). Row B, the same images with Z-score map overlays the minimum threshold set to 3 and maximum threshold set to 10. Red Circle: Incorrectly highlighted normal voxels. Blue Circles: Abnormalities core highlighted but edges missed.

pathological cases. Using this measure, the GP approach highlighted multiple types of tissue injury. This included haemorrhage, (e.g., intraventricular, cerebellar and germinal matrix), periventricular leukomalacia, punctate lesions and cysts (including sub-arachnoid and pseudocysts).

However, in areas of high inter-subject heterogeneity, most noticeably around the cortical surface, a number of false positives could be seen (Fig 3.2B3 red circle). This is likely due to inaccurate image registration between different cortical folding patterns across individuals and over the developmental period.

For the cross-validation of healthy controls, it was found that the false positive rate decreases from 0.013 to <0.001 when the Z threshold increases from 2 to 4.5, representing a reduction in the number of voxels being erroneously identified as abnormal, although given the large overall

number of voxels this is still large. The statistical results for the neonates with obvious lesions are shown in Table 3.3.

Table 3.3 The mean specificity, sensitivity and precision of models and a threshold of $Z > 3$.

The “standardisation” and “normalisation”, mean and median refer to the types of preprocessing each set of images had undergone. Values in bold indicate the best performance in their respective columns.

Test	Specificity	Sensitivity (Range)	Precision
Median standard	0.997	0.372 (0-0.967)	0.110
Median normal	0.999	0.049 (0-0.592)	0.065
Mean standard	0.997	0.372 (0-0.967)	0.110
Mean normal	0.999	0.049 (0-0.592)	0.065
Z-scores	0.997	0.291 (0-0.954)	0.095

The first notable result is that there was a clear effect on performance depending on the type of image pre-processing method used. Using a normalization process had a deteriorative effect on performance, resulting in lower scores than the raw Z-score segmentations. This included lower specificity and sensitivity scores across all Z-score thresholds.

In comparison, using within-image standardization before GP modelling did produce more accurate segmentation maps. For this reason, only the performance of the standardized input GP models will be examined in detail. Use of either the mean or median had no discernible effect for either the standardization or normalization practices.

In order to have a baseline measure of performance it is best to first look at the raw voxel-wise Z-score segmentation results, as some measure of improvement over these is desirable. Alongside this, equivalent scores will be given between the two manual segmentations to give an estimation of inter-rater variability. These are all shown in Table 3.4.

Table 3.4 Z-Score threshold comparison. The first row shows the mean inter-rater specificity, sensitivity and precision scores. The following rows show same mean scores for the raw Z-scores and median standardised image Z-score maps, both with thresholds of $Z > 2$, 3 and 4.5 applied.

Threshold/Inter-rater	Specificity	Sensitivity	Precision
Inter-rater	0.979	0.675	0.487
Raw $Z > 2$	0.898	0.465	0.029
Raw $Z > 3$	0.997	0.291	0.095
Raw $Z > 4.5$	0.999	0.196	0.156
Median standard $Z > 2$	0.989	0.533	0.036
Median standard $Z > 3$	0.997	0.372	0.110
Median standard $Z > 4.5$	0.999	0.204	0.205

These scores show an enhancement in performance by the GP based models over the raw Z-score segmentations and near inter-rater performance with higher specificity but lower sensitivity and precision for the Z threshold > 2 .

All models have high levels of specificity, only by lowering the SD threshold to 2 do values drop below 0.99. The sensitivity appears to be relatively low with only a threshold of 2 SD being able to detect over 50% of the abnormal voxels. Precision was mostly low, 0.110 for Median standard (best result) and rising to 0.241 when the SD threshold is raised to 4.5

representing a higher proportion of true positives. This low precision score of the Median standard $Z > 2$ is interesting as the increase in sensitivity, compared to the inter-rate, should indicate a large reduction in the number of false positives. The reduction in specificity does not appear enough to cause this drop in precision.

The produced ROC curves and PRC for each model including all tested methods as well as simple Z-score outlier detection for comparison are shown in Fig 3.3. The AUC for each of these ROC curves was also calculated and is shown in Table 3.5.

Table 3.5 Area under curve of receiver operating characteristic for each of the preprocess methods and raw Z-score against the individual raters. Values in bold indicate the best performance in their respective columns.

Pre-processing Type	AUC Rater1	AUC Rater2
Median standard	0.961	0.947
Median normal	0.927	0.907
Mean standard	0.961	0.947
Mean normal	0.927	0.907
Z-Score	0.934	0.915

Fig 3.3A shows the Median Standard ROC and Fig 3.3B the PRC for all individuals against both manual segmentations. Fig 3.3C and D shows the AUC score difference between simple Z-score outlier detection and the Median standard detection by age and by lesion size. Of intuitive note is the fact the blue GP model scores are consistently higher than their comparative orange raw Z-score values.

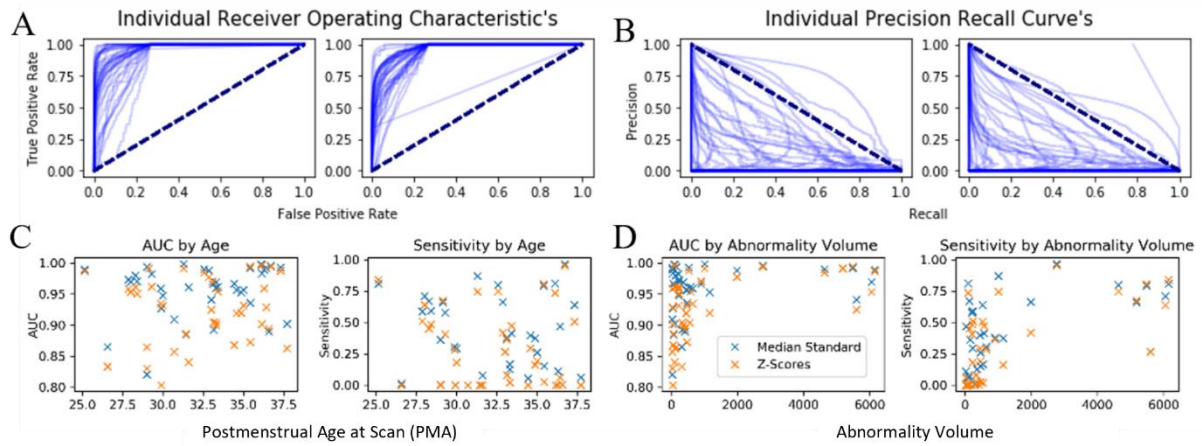


Figure 3.3 Individual pathology detection performance. A: Overlay of individual median standard ROC curves for rater 1 (left) and rater 2 (right). B: Overlay of individual median standard PRCs for rater 1 (left) and rater 2 (right). C: AUC score difference between raw Z-score outlier (Orange) and median standard (Blue) detection both at a SD threshold of 3 by age and D: by lesion size.

Wilcoxon p-scores were calculated to ascertain if the difference in model AUC performance between any two methods was statistically significant. As the mean and median feature scaling had no discernible effect, only three tests were performed. These were between the median standard model and the raw Z-score, the median normal model and the raw Z-score and the median standard and the median normal models. The median standardisation model showed a statistically significant higher AUC compared to both the raw Z scores and median normalisation model, attaining $p < 3.617e-5$ and $p < 1.196e-6$ respectively. The difference in AUC performance was not significant between median normal and raw z-scores in detection accuracy ($p < 0.295$).

3.4 Discussion

Voxel-wise Gaussian processes trained on healthy term-born and preterm neonates were used to estimate typical neonatal brain images appropriate to age and prematurity. These, in combination with generated variance maps and acquired neonatal MRI images, were used for

the purpose of segmenting heterogeneous brain tissue injuries in a held-out testing set of neonates with MRI detected pathology. The GP model performance was also quantified to sets of manual lesion segmentations. These resulting outlier (Z-score) maps can be overlaid on the original clinical T2w image, converted to absolute values, and the threshold changed to highlight areas of the brain believed to contain pathological tissue (Fig 3.2), acting as a form of computer aided pathology detection.

Unfortunately, Ghosh et al., 2014's work on HIE only described their mouse model outputs, having higher sensitivity but lower specificity. He & Parikh, 2013, Li et al., 2019, Bahadure et al., 2017, Pereira et al., 2016, Merzoug et al., 2023, Subudhi et al., 2020 and Rezaei et al., 2017 all presented segmentation methods, which showed good dice scores representative of higher balanced sensitivity and specificity compared to this work. Subbanna et al., 2019 stroke lesions detection performance was more in line with my own. One fact to bear in mind, is that these methods were all applied to a single lesion type, with more easily detectable features compared to some lesions within my own dataset (e.g., single punctate lesions).

Ramírez et al., 2018 demonstrated sensitivity and specificity scores similar to my own, while those seen in Richhariya et al., 2020 were higher. That said both were classification rather than segmentation methods, and as such, their results are less directly comparable.

An initial point that needs to be expanded upon is the sensitivity scores for the GP models. At first appearance it seems relatively low, however this is due both threshold selection and the effect of averaging over all the testing images. Individual sensitivity scores for the GP models ranged from 0.967 to 0 when a threshold of (absolute) $Z=3$ was applied to their Z-score maps. These represent the disparity in the level of injury present within the neonates MRI image. Sensitivity scores of 0, in which no pathology was successfully detected by the models, were recorded in 5 images with small or very subtle lesions bringing down accuracy, Fig 3.4. Images

with larger lesions had greater sensitivity scores and, in most of the remaining cases, the ‘core’ of the lesion was successfully highlighted.

In four of five images with sensitivity scores of 0 the neonatal MRI contained only one or two PWMLs and the 5th was a cerebellar haemorrhage. All were very small lesions with subtle visual differences between abnormal and normal tissue intensity, Fig 3.4. This made pathology both challenging to detect and unlikely to have a permanent effect on the neonate’s outcome (Carney et al., 2021; Kumpulainen et al., 2020). I did not exclude these values from calculations as they reflect true testing conditions and clinical observations. Although the core of larger lesions is usually segmented correctly, there can sometimes be insensitivity in the extent of the lesions, and it is here that the GP models can sometimes underestimate the volume by missing the outer edges. This is likely because the intensity values at the lesion edges are closer to normal tissue values and the boundary of the pathology can be ambiguous.

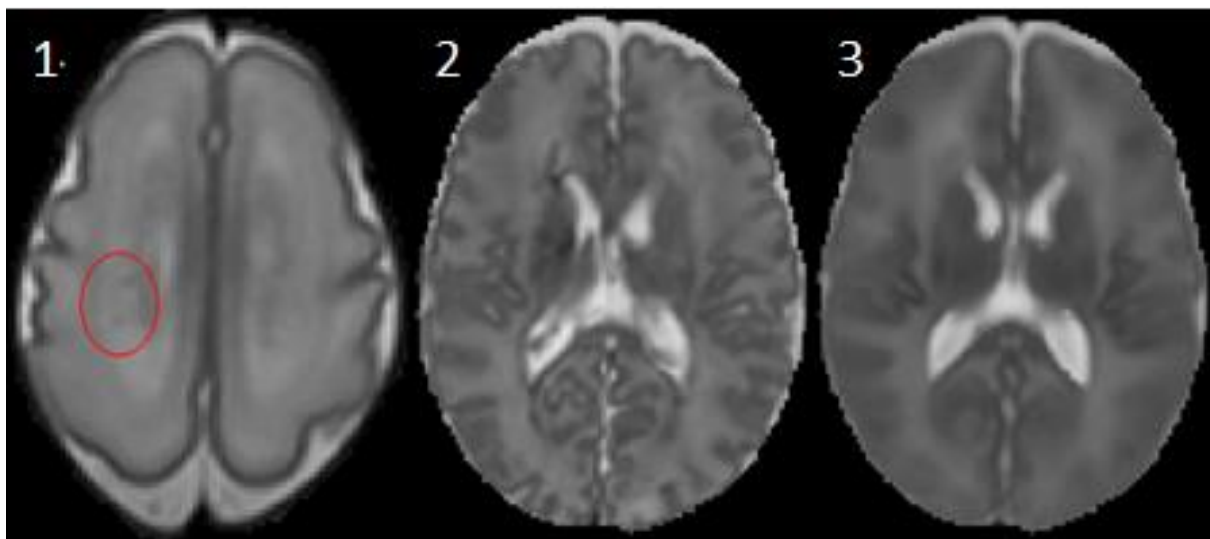


Figure 3.4 Missed pathologies and smoothing of the predicted images. Column 1 shows two almost indistinguishable punctate white matter lesions inside the red circle. Columns 2 and 3 show the raw and GP predicted images. Note how the cortical surface is very smooth around the cortical surface as the model needs to accommodate multiple tissue types.

Specificity remains high since the number of voxels in an MRI image is extremely large ($\approx 90,000$). Even if an image is noisy, with 100 falsely segmented voxels for example, it is still relatively small compared to the total number of voxels which overpowers their effect (H. He & Garcia, 2009; Varoquaux, 2018). This is why a threshold of 3 was decided as most logical as it was the point at which, visually, random noisy voxels almost completely disappeared from the segmentation. Data driven thresholds would be a future enhancement.

The highly variable nature of the cortical surface resulted in several false positives in these areas when CSF is present in the subject image (Fig 3.2C red circle) and is primary contributor to the low precision. The cortex was blurred in these areas and larger GP variance output values were found as the model needs to account for the range of intensities that may be present in these voxels, Fig 3.4 2-3.

One potential option to improve lesion coverage would be lowering the Z threshold but at the cost of a higher number of false positives. However, this may not necessarily be a bad thing depending on exactly how the method is being utilized or if some form of noise suppression could be incorporated such as that seen in Saladi & Amutha Prabha, 2018, Coupé et al., 2018, Szabó et al., 2018 and Subbanna et al., 2019. For quick identification of critical patients, those most likely to have a serious injury with a notable intensity difference in the MRI, a high threshold should be suitable. If looking for more subtle injuries or trying to estimate lesion extent or volume a lower threshold would be better but may need additional steps to account for the extra noise. Alternatively, some form of secondary method could be used to decide if there is a lesion present within an MRI based on the Z-score image such as seen in Subudhi et al., 2020.

A point that needs to be reemphasised is the generalizability of this method. Most of the reviewed abnormality detection methods focus on detection only a single pathology. As this

method utilizes the deviation from an expected normal value, it is invariant to the type of pathology itself. It is reliant only on the requirement that there is some form of intensity difference associated with the pathology's MRI appearance.

This method could be easily expanded to include other modalities such as T1w images and diffusion-based images (e.g., apparent diffusion coefficient (ADC)) as multi-modal imaging is recommended to improve diagnostic accuracy (Miller et al., 2020; Rutherford et al., 2010; Volpe, 2012). This may be done either using separate sets of models each of which is specific to an MRI modality or alternatively as part of a multi-output GP.

An enhancement that can be implemented into the GP models that may improve its ability to match individual brain structure is the incorporation of neighbourhood information via a patch surrounding the predicted voxel (Guizard et al., 2015; Roy et al., 2018; N. Subbanna et al., 2015). This will allow the model to infer additional spatial information specific to an individual. It is hoped that this will make prediction of each voxel, and therefore the predicted normal subject image, more specific to each individual. One area where this will be of particular use is the cortical surface where it can assist in identification of tissue boundaries and reduce blurring. A caveat is that care will have to be taken to ensure that this does not cause overfitting.

3.5 Conclusion

The performance of the GP models was shown to significantly outperform baseline raw Z-scores. Depending on the choice of threshold applied to the absolute Z-score map, this method showed similar sensitivity and specificity values as seen when comparing two manual segmentations from two separate raters.

The voxel-wise Gaussian Processes normative modelling method proposed in this chapter can model typical intensity values in a T2w modality image of the brain throughout the neonatal

period. These predicted images, along with the corresponding variance maps, can be used to generate Z-score abnormality maps that give a voxel-wise estimation of intensity deviation a neonates MRI has compared to typical tissue intensities. By applying a threshold to these Z-score maps, it is possible to detect voxels whose tissue intensity values deviate to a level that they are likely to represent lesions and abnormalities. This provides an automated approach to detection of a wide range of abnormalities which can be observed within the neonatal brain.

Chapter 4. Detection of Injury and Automated Triage of Preterm Neonatal MRI using Patch-Based Gaussian Processes

Automatic detection or highlighting of neonatal brain injury could be a valuable adjunct to radiological interpretation. Here I propose refinement to the normative modelling-based detection method for preterm neonatal neuroimaging using gaussian processes (GPs) seen in the previous chapter. This includes the incorporation of local anatomical information from patches of voxels from a subject MRI image and demographics such as age and sex. Z-score images can then be created from the scaled difference between the model predictions and a neonate's T1 and T2 weighted MRI. It also examines the use of these GP Z-scores as a form of automated triage where a logistic regression classifier is trained to separate normal and abnormal images. 133 preterm neonatal images with normal-reported MRI are used to train a GP model and lesion detection performance optimized on 36 preterm neonatal images with manually annotated lesion masks and tested on 20 neonates with brain injury. The automated triage model was trained on 100 preterm neonates with normal reported MRI and 109 preterm neonates with MRI detectable lesions. It was tested on the same 36 manually annotated abnormal MRI preterm neonates and 33 normal-reported preterm neonates. A second, more robust test was performed using 5-fold cross-validation on 73 neonates from the brain injury cohort and 73 randomly selected neonates with normal appearing MRI. Using a patch diameter of 7 voxels and integrating both T1w and T2w Z-score images provided our highest performing GP model for within image lesion detection, achieving an AUC of 0.961. The best classification accuracy was achieved by combining the output probabilities of separate T1w and T2w Z-score histogram classifiers. This allowed for correct identification of 32/36 abnormal and 28/33 normal images from the preterm cohort and 64/73 abnormal and 62/73 normal images in the held-out brain injury cohort. These results indicate a patch-based normative model can generate abnormality maps to accurately detect lesions in neonatal MRI in a highly interpretable fashion

and, at a group level, can be used to classify a group of neonates with MRI into those who have typical and atypical brain structure.

4.1 Introduction

As highlighted in the preceding chapters, the neonatal brain is a complex structure, with variability in shape and volume between individuals and across age (Groeschel et al., 2010). This complexity is amplified by the various developmental processes, such as gyrification, myelination and volumetric growth, which progress at a rapid pace over the first few months of life (Miller et al., 2020). This combination of factors makes identification of abnormal tissue, structure and pathology challenging, requiring expert knowledge of typical developmental brain anatomy, image contrast changes and the appearances of injuries themselves.

Given these developmental factors and the effect their disruption or brain injury can have on the long-term outcome of a neonate, efficient detection is crucial (Anand & Scalzo, 2000). However, the increasing use of MRI for assessment of neonatal neurological injury combined with the extensive assessment times necessary for review, mean that diagnosis could be longer than would be desirable. This could lead to a situation where neonates with more severe injuries and large lesion loads are not prioritised over neonates with only suspected injury and normal appearing MRI with little diagnostic use. Some method of triage that can isolate these more injured cases and prioritise their review would be highly desirable.

The voxel-wise models in the previous chapter provided predictions that were spatially smooth, as models were specific to the time-point being modelled and did not consider neighbouring tissue structure. This made them sensitive to local shifts in intensity, typical brain shape variability, and misregistration errors. In predicted images, highly variable areas (such as the cortical surface) and structural boundaries appeared blurry, reducing interpretability. These effects are due to tissue type variability across the sample (white matter (WM), grey matter

(GM), and cerebrospinal fluid (CSF)) and differences in exact position (brain structure and cortical surface heterogeneity, registration performance). This can result in areas of rare, but still typical, anatomical variants being classified as abnormal (false positives) and areas of actual abnormal tissue intensity falling within normal tissue variability and not being detected (false negatives). This is an issue affecting other methods that base detection of MRI lesions on differences from expected norms. A recent review by Baur et al., 2021 showed that synthetic images created by generator adversarial networks for the purpose of lesion detection showed similar issues of overly smoothed images or over inferencing from the normative cohort.

Ideally, a predictive model of anatomy should incorporate normal local tissue heterogeneity into the prediction while inferring typical tissue intensities from the training set in abnormal areas. Prior work has utilised a variety of methods to incorporate local spatial information into models. Subbanna et al., 2015 used the graphical nature of Markov random fields to incorporate surrounding voxel intensities when making tissue segmentations. Essa et al., 2020 and Roy et al., 2018 both used adaptations of convolutional neural networks, which natively incorporate local patch information, for the detection of multiple sclerosis lesions. Kaplan et al., 2022 used nested patches to extract texture features to identify Parkinson's disease patients using k nearest neighbour and support vector machine classifiers. A particularly interesting case is Guizard et al., 2015 which used a model similar to my own but made comparisons between patches rather than individual voxels. This is of interest as an adaption of this method could be used to enhance the model created in the previous chapter.

In the previous chapter, each GP model made predictions based purely on age (gestational age at birth (GA), post menstrual age at scan (PMA)) and sex. No information about the local tissue, and therefore any individual heterogeneity, was incorporated into the model. An enhancement could be inclusion of intensity values from the surrounding voxels into the design matrix

alongside this demographic information. When making predictions using a voxel-wise GP model, this would be equivocal to comparing patches of tissue around the voxel to be predicted.

The proposal for this chapter is to expand voxel-wise GP models by incorporating subject-specific local structural information in the form of patches of voxels into the model. The aim is to increase spatial accuracy of the GP models by incorporating local anatomical and image heterogeneity into the predictions. This will hopefully lead to predicted images that are sharper and that more closely match an individual's actual MRI, making them more interpretable. It could also potentially improve abnormality detection via reduction of false positives in highly variable areas.

Using the outlier maps created by the patch-wise GP method, I will also develop a classifier to separate neonates with healthy appearing brains from those containing obvious pathological tissue. This essentially acts as a form of triage, indicating which neonates have abnormal MRI for prioritisation for radiological follow-up. To better estimate the accuracy of the within subject lesions detection and the across subject classification of pathological neonates using the voxel-wise GP Z-score maps, application to a second brain injury cohort with similar acquisition parameters was performed.

4.2 Methods

4.2.1 Sample and dataset

Ethical permission for this study was granted by the Hammersmith and Queen Charlotte's and Chelsea Hospital (QCCH) Research Ethics Committee. Written parental consent was obtained for each infant. The same preterm MRI dataset of 423 used in chapter 3, acquired on a Philips 3T scanner, were used for modelling and testing. Modelling was implemented on some form of T1w image, high-res images when available but non-high-res images otherwise, and T2w images (Table 4.1). After quality control, 145 were discarded due to not having both a T2w

and some form of T1w image, either standard T1w or High Res T1w, of sufficiently high quality, based on the same exclusion criteria seen in chapter 3, leaving 278 complete datasets.

Table 4.1 Acquisition parameters.

Scan	TR	TE	Voxel Size
T2w	7000-8000ms	160ms	0.859 x 0.859 x 1mm
T1w	17ms	4.6ms	0.8 x 0.82 x 0.82mm
High Res T1w	17ms	4.6ms	0.5 x 0.82 x 0.82mm

Of the 278 images, 133 were considered “normal” in a preterm context and used as training data for the GP models as they contained no obvious lesions and the remaining 145 used for testing. Of the 39 manually labelled images from chapter 3, only 36 had an additional T1w image required for this chapter of the thesis. They included the following pathologies: Germinal matrix haemorrhage (n=11), intra-ventricular haemorrhage (IVH) (n=6), cerebellar haemorrhage (n=11), haemorrhagic parenchymal infarct (n=4), temporal horn cysts (n=1), sub-arachnoid cysts (n=2), pseudocysts (n=1), cystic periventricular leukomalacia (n=3), large(>8 voxels)/many(>3) punctate white matter lesions (PWML) (N=7) and small (<=8 voxels)/few(<=3) PWML (N=9). 13 neonates had multiple pathologies.

The second brain injury cohort is composed of an additional 73 pathological neonates, 20 of which included manual segmentations of the areas. The injuries found in this cohort include stroke (n=2), hypoxic ischemic encephalopathy (HIE) (n=3), enlarged ventricles /ventriculomegaly (n=2), meningitis (n=2), PWML (n=2), infarct (n=2), diffuse excessive high signal intensity (DEHSI) (n=1), IVH (n=1), cerebellar lesions (n=1), WM lesions (n=5), brain stem lesions (n=1), venous thrombosis (n=1), agenesis of the corpus callosum (n=1), posterior subcortical WM haemorrhage (n=1) again with 6 neonates having some overlap.

Subject demographics are displayed in Table 4.2. Manual segmentations were created by 2 raters using T2w images and inter-rater reliability assessed as opinion on lesion extent can vary (Bogner, 2018).

Table 4.2 Cohort Information.

Parameter	Value
Total Preterm Subjects (Female) [High Resolution T1 Image]	278 (125) [45]
Post Menstrual Age at scan (PMA) Median (Weeks+Days) [Range]	39+1 [25+1 to 55+0]
Gestational Age at birth (GA) Median [Range]	29+4 [23+2 to 42+0]
Preterm without Lesions – Normative Model Training	133 (66) [30]
PMA Median [Range]	40+4 [27+5 to 51+0]
GA Median [Range]	30+0 [23+2 to 35+4]
Preterm with Lesions – Primary Testing Cohort	145 (59) [13]
PMA Median [Range]	33+2 [25+1 to 55+0]
GA Median [Range]	29+3 [24+3 to 36+1]
Preterm with Lesions and Ground Truth - Statistical Testing	36 (18) [2]
PMA Median [Range]	33+0 [25+1 to 37+5]
GA Median [Range]	30+3 [24+3 to 35+2]
Brain Injury Cohort - Replication Testing Cohort	73 (39) [73]
PMA Medium [Range]	40+4 [28+2 to 46+3]
GA Median [Range]	38+3 [24+3 to 42+1]
Brain Injury Cohort with Ground Truth - Statistical Testing	20 (12) [20]
PMA Median [Range]	40+0 [32+3 to 45+4]
GA Mean [Range]	38+0 [24+3 to 41+1]

4.2.2 Pre-processing

Individual subject T1w images were initially rigidly co-registered using FSL (Jenkinson et al., 2002; Jenkinson & Smith, 2001) to their equivalent T2w images. B0 inhomogeneity was corrected using N4BiasFieldCorrection (Avants et al., 2014) followed by removal of non-brain tissue using FSLBET (Smith, 2002). Affine registrations between the T2w images and weekly developing human connectome project templates (Makropoulos et al., 2018) were calculated using FSL (Jenkinson & Smith, 2001), followed by a non-linear registration to the same template using ANTS (Avants et al., 2014). Datasets below 28 weeks PMA were registered to week 28, the youngest template available. Each weekly template image also had an affine registration to the week 36 template created which was applied to the images to bring them into spatial and size alignment. These transformations and warps were applied to the T1w images to bring them into alignment with the T2w images within template space. Finally, all voxels within an image were standardized to image intensity median and standard deviation.

4.2.3 Base Gaussian Process Tissue Intensity Model

A separate GP model was fit to each voxel (Rasmussen & Williams, 2005), estimated using GPyTorch (Gardner et al., 2018). GPs model a distribution of possible functions that fit some dependent variable given an input dataset. A prior ε (an initial estimation of the intensity distribution of voxel n) is specified based on a normal distribution N with a mean (usually 0) and variance σ^2 to constrain the distribution, formula 4.1.

$$\varepsilon \sim N(0, \sigma_n^2)$$

4.1

The covariance between datapoints is quantified by the widely used radial basis function (RBF), formula 4.2, as a covariance matrix Σ , formula 4.3.

$$K(x, x) = \sigma^2 \exp\left(-\frac{d(x, x')}{2l^2}\right)$$

4.2

$$\Sigma = K(x, x)$$

4.3

The RBF kernel hyperparameters (length-scale and variance) were optimized using the Adam optimizer (Kingma & Ba, 2014) which minimizes the log marginal likelihood. Predictions were made based on the joint conditional probability of the training and testing samples, formula 4.4.

$$\begin{bmatrix} f \\ f^* \end{bmatrix} \sim N\left(0, \begin{bmatrix} k(x, x) & k(x, x^*) \\ k(x^*, x) & k(x^*, x^*) \end{bmatrix}\right)$$

4.4

Here, each voxel's GP incorporated GA at birth, PMA at scan and sex in the design matrix to model image intensity. This represents the current state of the model after creation in the previous chapter and is henceforth referred to as the base model.

4.2.4 Patch Extension to the Gaussian Process Tissue Intensity Model

Patch-based models used the same structure as the base model but incorporated additional spatial information at training and inference time. Intensity values from a patch of voxels surrounding the target voxel were added into the design matrix. This provides a prior on target voxel intensity from surrounding patch voxels making the model less dependent on global image standardisation. It also allows comparison of intensity values within a patch of the neonate being predicted and those seen in the normative cohort. Voxel GPs with similar patches will make predictions with similar outputs but those with patches that differ greatly from those seen in normal cohort will instead revert to the posterior mean prediction.

Different sizes of patch neighbourhoods of 3x3x3, 5x5x5, 7x7x7 and 9x9x9 using two approaches of voxel contribution were investigated. These two approaches were a dense patch, which used all voxels except the target voxel for prediction, and a surface patch, that only used those voxels on the outermost edge of the patch. An example of how this is implemented is demonstrated in Fig 4.1 for a 5x5x5 patch.

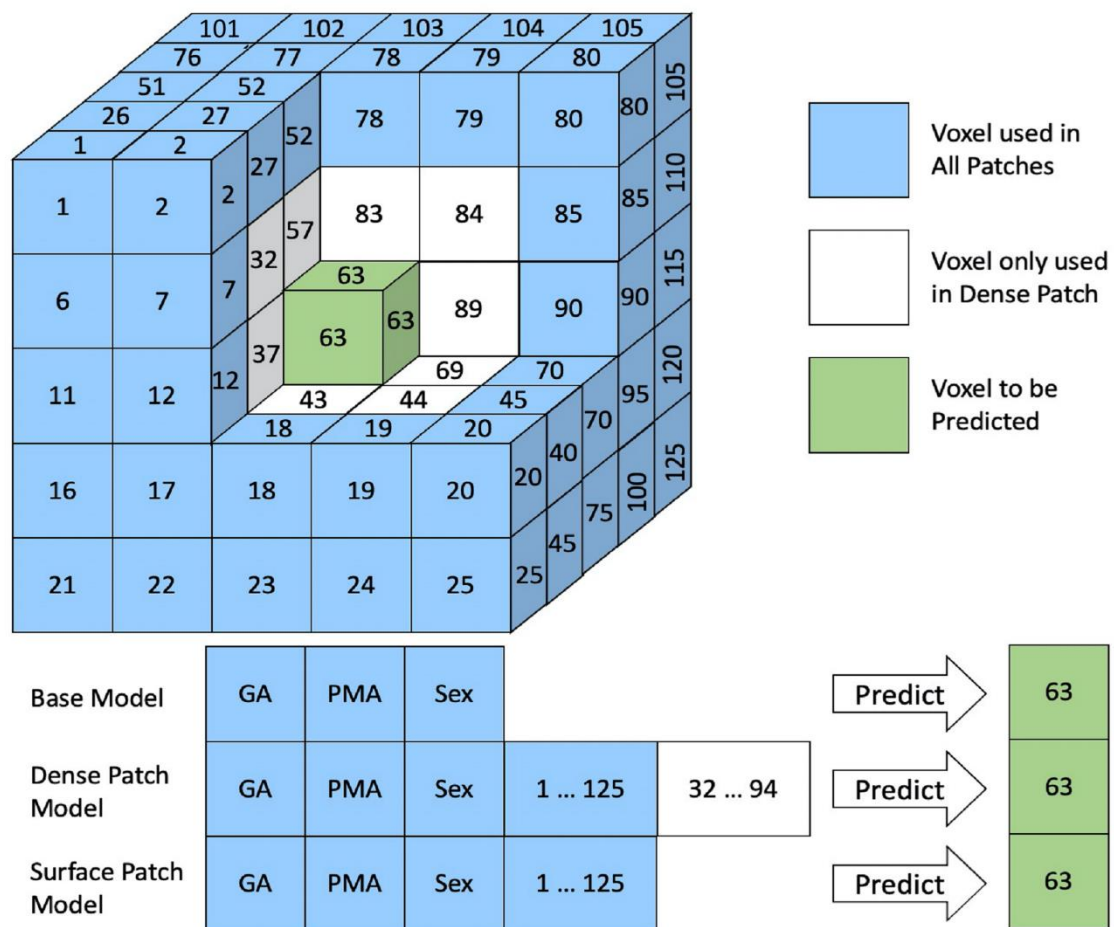


Figure 4.1 Example 5x5x5 patch model. A model predicts the intensity of the green voxel based on the gestational age at birth (GA), post menstrual age at scan (PMA), sex, and any patch voxel intensities included in the design matrix. For this 5x5x5 patch, the dense patch included white and blue voxel intensities while the surface patch included only blue voxel intensities.

4.2.4.1 Iterative Patch Extension Gaussian Process Tissue Intensity Model

Lesions, when present within a patch, have voxel intensity values outside the range observed within the normative (training) cohort. This leads to the model being less certain of its prediction and outputting a high predictive variance, even if the posterior based mean prediction itself is a reasonable estimation of the target voxel intensity. This would have the effect of reducing the Z-score values in the areas of abnormality and potentially masking them during the threshold process, the antithesis of what this project aims to achieve.

To address this, the effect of iteratively applying the model by replacing the actual MRI image patch input with patches extracted from the model's predicted image was investigated. This acts as a form of lesion infilling by replacing outlier tissue intensities that may have been present in the actual MRI image with normal values estimated by the model. This iteration process was applied five times with the initial model using the actual MRI and all subsequent models using the output of the previous model. This effect of iterating model was also applied to each of the patch sizes and type (dense and surface).

4.2.5 Detecting Pathology

Z-score images were created by calculating the difference between the observed and predicted images and scaling by the standard deviation derived from the output variance. Receiver operating characteristic's (ROC) area under the curve (AUC) were calculated for the 36 held-out images from the initial preterm brain injury cohort and the 20 neonates from the second brain injury cohort both with manual lesion labels. This was calculated between each neonate's absolute Z-score image with fixed thresholds applied against their corresponding manual segmentations and averaged across subjects for each model. Separate Z-score maps were made for the T1w and T2w images. A third was created by finding the mean of the T1w and T2w absolute Z-score maps.

For the best performing Z-score maps (T1w, T2w or mean T1w and T2w), sensitivity and specificity scores were calculated at a series of fixed thresholds (two, three, four and five). These provided a more precise estimation of model pathological detection compared to AUC as the task suffers from imbalance issues due to the unbalanced nature of the task. Specifically, as there are far more voxels with normal intensities than abnormal ones even in pathological neonates, specificity is always likely to be high, even in noisy predictions with many false positives.

4.2.6 Automated Clinical Image Triage

The logistic regression classifier was selected to separate images into those with normal appearing brains (1) or those with lesions (0) due to its simple implementation and high interpretability. Histograms, contained 99 equally spaced bins between -100 and 100, were extracted from the T1w and T2w Z-score maps for use as inputs to two classifiers. A third classifier was constructed based on the mean of the T1w and T2w absolute Z-score maps because, as noted in the introduction, different modalities often contain complementary information (Rutherford et al., 2010).

Alongside these three one-stage classifiers, a fourth two-stage classifier was investigated based on the logistic regression models probabilistic outputs. Rather than the binary (0,1) output, the probabilistic score (between 0 and 1) of the separate T1w and T2w classifier predictions being combined via mean was investigated. For example, if a pathology has an obvious appearance in T1w modality images it may have a high predictive value in that model but perhaps not in the equivalent T2w model, while the opposite may be true in a second type of pathology. By combining the predicted probability of the two models it may be possible to reach the ideal situation whether mean of both models is high enough to give the correct classification.

As an example, imagine two neonates with different MRI positive pathologies each of which is clearer on separate modality types. The T1w model gives results of 0.86 and 0.34 while the T2w model gives 0.42 and 0.73. Both correctly identify one neonate each and both miss one neonate. Finding the mean of the predictions gives the values 0.64 and 0.535, which after having the 0.5 threshold applied give two positive outputs, correctly identifying both neonates.

In order to create predicted images of the normal appearing neonates, 5-fold cross application of the model with the best AUC from the voxel-wise segmentation analysis was applied to the 133 preterm neonates without obvious lesions for three iterations. Of these, 33, stratified by age, were added to the 36 manually segmented images with abnormalities to act as a classifier test set. The remaining 100 preterm images with no obvious lesions and 109 preterm images with lesions were used for training. The number of correctly classified images, sensitivity and specificity were recorded.

To assess the model performance in a more robust manner, a more comprehensive analysis was completed using the normal appearing brains from the first dataset and the separate brain injury dataset. In this instance, the 73 neonates from the brain injury cohort and an equal number randomly selected from the neonates without lesions were combined and a 5-fold cross validation of the classification models applied.

4.3 Results

4.3.1 Base Gaussian Process Tissue Intensity Model

The synthesized images of the base model offered a good qualitative representation of neonatal brain structure but as expected, were very smooth. With the exception of major sulci and gyri, heterogeneous cortical folding patterns (as in frontal and parietal areas) were not modelled well, appearing blurry. Nonetheless, for detecting lesions, the base model had a high joint AUC

of 0.948, a T2w AUC of 0.951 and the T1w model performance being slightly lower with an AUC of 0.918, Table 4.3.

Table 4.3 AUC for all models on the 1st and 3rd iteration. Values in bold indicate the best performance in their respective rows.

Model - Iteration	Base	Dense 3x3x3	Dense 5x5x5	Dense 7x7x7	Dense 9x9x9	Surface 5x5x5	Surface 7x7x7	Surface 9x9x9
T1w – 1	0.918	0.901	0.919	0.919	0.929	0.926	0.937	0.938
T2w – 1	0.951	0.904	0.92	0.92	0.928	0.927	0.937	0.938
Mean of T1 and T2 - 1	0.948	0.915	0.924	0.924	0.942	0.935	0.951	0.953
T1w – 3		0.908	0.932	0.941	0.941	0.941	0.947	0.946
T2w – 3		0.911	0.934	0.943	0.942	0.943	0.948	0.946
Mean of T1 and T2 - 3		0.925	0.94	0.953	0.952	0.953	0.961	0.959

4.3.2 Patch Based Gaussian Process Tissue Intensity Model

Synthesized images produced using patch information incorporated more subject specific structural heterogeneity compared to the base model but also included some pathological

hypo/hyperintensity into the predicted image, Fig 4.2. Patch size affected this generalisation, smaller patches closely matched heterogeneous brain structure but included greater amounts of local pathology, while larger patches were smoother but included less abnormal tissue intensities. A similar effect was apparent in the dense and surfaced patch models, dense patches being better at approximating individual heterogeneity at the cost of more abnormal tissue intensity inclusion when compared to models utilising surface patches.

GPs make predictions via a similarity measure and, in most of the brain, target voxel patches were consistent with those seen in the training set. This has the effect of greatly improving model confidence, resulting in large reductions to the model variance output values to within the range of 0.003 to 0.006. In lesioned areas, patches could contain intensity values drastically different from those in the training set. This reduced model certainty and dramatically increased output variance (≈ 0.05 , an order of magnitude increase) compared to areas of uninjured tissue. The predicted mean output in these lesioned areas reverts to the optimised posterior mean (the equivalent to the mean of all intensities for a given voxel). As the true underlying structure of these areas is obscured by the pathology, these more general predictions still represent a good estimation of normal tissue. This results in a reasonable predicted image but high variance values in areas of abnormality. This leads to Z-scores in abnormal area having values like those seen in healthy tissue (Fig 4.2. Yellow circle).

This effect could be alleviated to a degree by using larger surface-based patches. Fewer voxels, further away from the target voxel both reduced the chance an abnormal voxel would be present in the patch and the model's ability to infer intensity values for a given input. Synthesis of "normal" brain tissue in surface patches remained similar to the dense patch models in all but the 9x9x9 patch which contained a more blurring in the cortical surface.

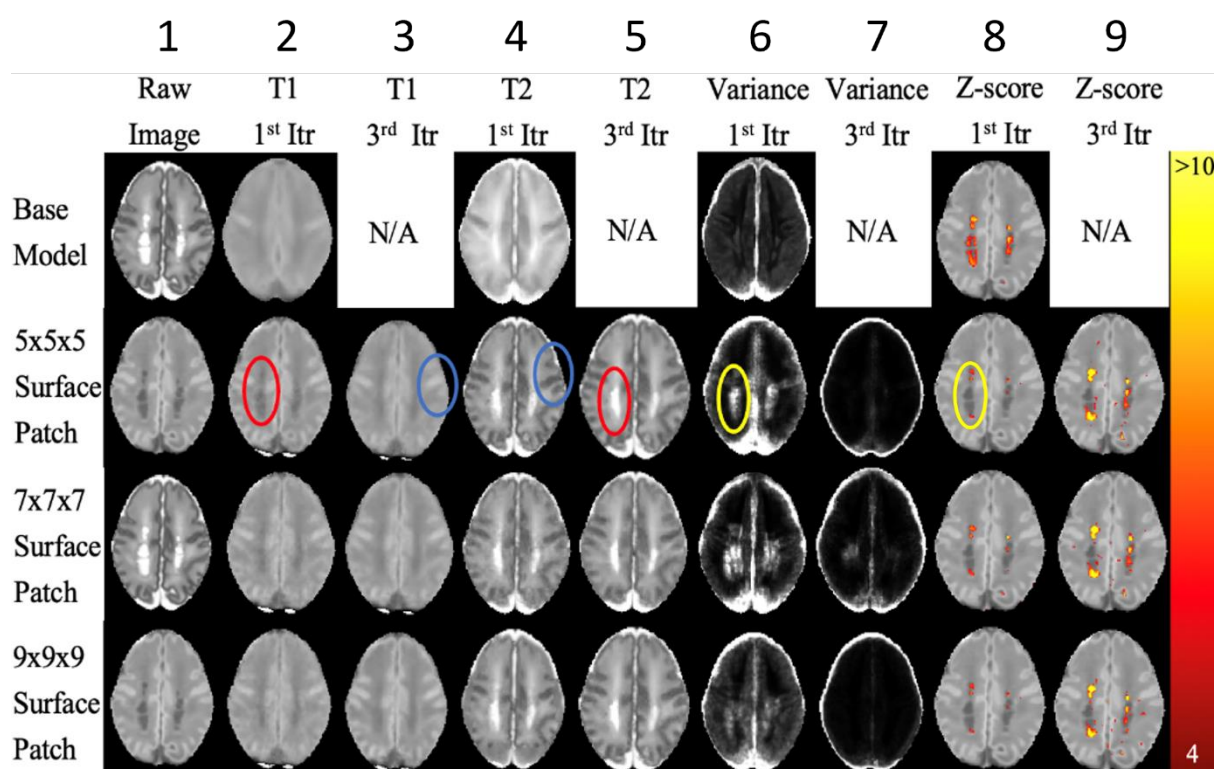


Figure 4.2 T1w/T2w predicted images, T2w variance maps and absolute T2w Z-scores maps.

Column 1 shows the raw T1w and T2w images, columns 2 and 3 the T1w images after a single and 3 iterations, and columns 4 and 5 the T2w images after a single and 3 iterations. Columns 6 and 7 show the T1w GP variance images after a single and 3 iterations. Columns 8 and 9 show T2w Z-score maps with minimum threshold set to 4 and maximum threshold set to 10 for a single and 3 iterations. Blue circles show heterogenous cortical structure that the patch models were able to duplicate. Red circles show lesion areas retained by the patch model. Yellow circles show associated high variance in patch models and its reduction via iteration.

The 9x9x9 surface patch using the mean of the T1w and T2w absolute Z-scores achieved the best AUC after a single iteration with a score of 0.953, compared to the base model's highest score of 0.951 using T2w absolute Z-score, Table 4.3. The patch based T2w AUC was lower than the base model. This was due to lower scores in neonates with larger lesions but was compensated by better coverage of small lesions. This loss of large lesion coverage was caused by the partial inclusion of the lesion cores into the predicted image and resulting in them not

being detected. The T1w model AUC scores were consistently lower than the T2w AUC score but averaging (mean) the T1w and T2w model output together gave the highest AUC scores for all patch-based models.

On the surface it would appear that the best patch-based model has not achieved a much greater performance than the best base model, however this is due to the averaging effect over multiple individuals missing an important distinction. In the base model, larger lesions have a more complete coverage compared to patch-based models due to some abnormal tissue intensity being included in the prediction. This did not result in a drastic loss of AUC because the smaller lesions, which the base model missed or underestimated, were detected more completely and consistently.

Iteration of the patch models by extracting intensity values from the previous iterations predicted images rather than the actual MRI image, resulted in higher AUC scores. However, after three iterations most models demonstrated no further improvement and subsequent iterations began to degrade the synthesised image. The 7x7x7 surface patch model which averaged (mean) the Tw1 and T2w absolute Z-scores after three iterations had the best overall performance and achieved an AUC of 0.961, Table 4.3.

Table 4.4 shows the sensitivity and specificity scores for models with AUCs above 0.95 after three iterations for a range of thresholds (two, three, four and five). For absolute Z-score thresholds of 2 to 4 the 7x7x7 surface patch model gave the best sensitivity, specificity trade-off but the 7x7x7 dense patch model had higher sensitivity for higher thresholds. This effect was mostly replicated when applied to the 20 neonates with manual segmentations found in the separate brain injury data set. The only major difference being that the surface 5x5x5 model had slightly better performance than all others at an absolute Z-score threshold of two, Table 4.5.

Table 4.4 Sensitivity and specificity for the 36 images with manual segmentations from the initial preterm dataset. Meaned T1w and T2w absolute Z-score models with AUCs higher than 0.95 after 3 iterations for absolute Z-score thresholds ranging from 2 to 5. Values in bold indicate the best performance in their respective paired sensitivity and specificity columns.

Z Thresh (Sens/Spec)	2 (Sens)	2 (Spec)	3 (Sens)	3 (Spec)	4 (Sens)	4 (Spec)	5 (Sens)	5 Spec)
Base	0.358	0.993	0.198	0.998	0.108	0.999	0.056	0.999
Dense 7x7x7	0.656	0.972	0.467	0.993	0.319	0.998	0.232	0.999
Dense 9x9x9	0.632	0.971	0.436	0.992	0.308	0.997	0.222	0.999
Surface 5x5x5	0.641	0.98	0.451	0.995	0.306	0.998	0.211	0.999
Surface 7x7x7	0.645	0.982	0.463	0.995	0.321	0.998	0.223	0.999
Surface 9x9x9	0.629	0.98	0.445	0.995	0.308	0.998	0.213	0.999

4.3.3 Automated Triage of Abnormal Images

The 7x7x7 surface patch model output was used as data for the triage model as it consistently demonstrated higher AUC, sensitivity and specificity scores for lesion detection. The best performance was seen by combining the predictive output of separate T1w and T2w histogram models, correctly identifying 32 of 36 abnormal images and 28 of 33 normal images and achieving a sensitivity score of 0.889 and a specificity score of 0.848.

Interestingly, the mean of the T1w and T2w absolute Z-scores gave the worst performance, only identifying 27 of 36 abnormal images and 23 of 33 normal images giving a sensitivity

score of 0.750 and a specificity score of 0.670. The full results of these tests can be seen in Table 4.6 and the corresponding results from the brain injury cohort in Table 4.7.

Table 4.5 Sensitivity and specificity for the 20 images with manual segmentations from the second brain injury dataset. Meaned T1w and T2w absolute Z-score models with AUCs higher than 0.95 after 3 iterations for absolute Z-score thresholds ranging from 2 to 5. Values in bold indicate the best performance in their respective paired sensitivity and specificity columns.

Z Thresh (Sens/Spec)	2 (Sens)	2 (Spec)	3 (Sens)	3 (Spec)	4 (Sens)	4 (Spec)	5 (Sens)	5 Spec)
Base	0.342	0.994	0.200	0.997	0.100	0.998	0.049	0.999
Dense 7x7x7	0.661	0.972	0.447	0.994	0.319	0.998	0.233	0.999
Dense 9x9x9	0.641	0.972	0.435	0.992	0.310	0.998	0.224	0.999
Surface 5x5x5	0.662	0.981	0.449	0.996	0.309	0.998	0.213	0.999
Surface 7x7x7	0.665	0.979	0.465	0.995	0.325	0.998	0.229	0.999
Surface 9x9x9	0.630	0.98	0.442	0.995	0.310	0.998	0.218	0.999

4.4 Discussion

In this work, I extended normative voxel-wise GP models to include local structural information into the model. By incorporating voxel intensities in a patch surrounding the target voxel, the GP model could synthesise typical MRI images of neonates more closely matching the neonate's actual anatomy. Different variations of patch size and contribution were examined in order to evaluate their effect on the model's predictive ability. The ability of the new patch-based model to qualitatively encapsulate individual anatomy while maintaining

good lesion detection performance was demonstrated. Finally, a logistic regression classifier based on the histogram of the Z-score maps created from GP outputs was used to separate abnormal and normal-appearing images, essentially acting as a form of triage. When combined, these two methods could allow a 2-stage strategy of investigation. First, automated triage to locate neonates within a cohort with MRI positive lesions that may need follow-up or special attention, followed by within image localisation of these lesions. If implemented within a clinical setting, these methods could help streamline and assist the radiological review process.

Table 4.6 Neonates from the initial preterm cohort that were correctly identified by the different 7x7x7 surface patch models. Rows 1-3 show models created from the T1w Z-score histograms, T2w Z-score histograms and mean of the T1w and T2w Z-score histograms. The 4th row shows the mean of the probabilistic output from the separate T1w and T2w Z-score histogram models. Each cell in the Abnormal and Normal Image columns shows the number of correctly identified images followed by the total number of images of that type. Values in bold indicate the best overall performance.

Model	Abnormal Images	Normal Images	Sensitivity	Specificity
T1w Z-score	28/36 (77.77%)	28/33 (84.84%)	0.778	0.848
T2w Z-score	30/36 (83.33%)	27/33 (81.81%)	0.833	0.818
Mean T1w and T2w Z-scores	27/36 (75%)	23/33 (69.69%)	0.750	0.670
Mean T1w prediction and T2w prediction	32/36 (88.88%)	28/33 (84.84%)	0.889	0.848

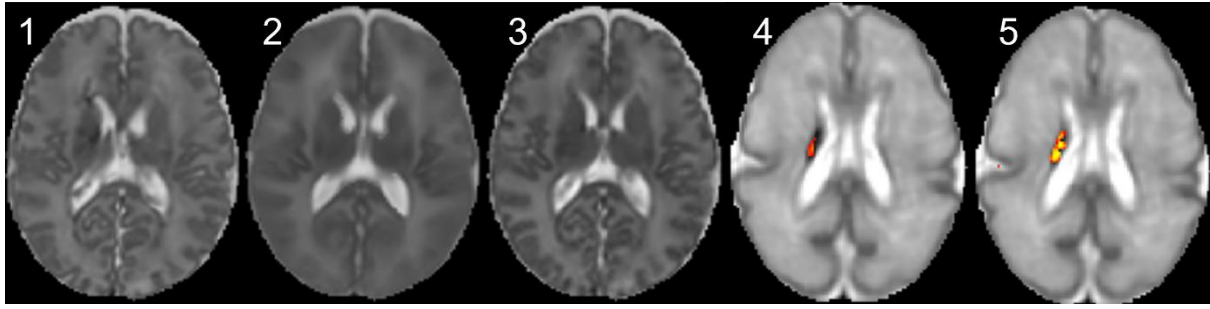


Figure 4.3 Improved patch-based model performance over the base model. Image 1 shows a subject's raw image, 2 the overly smoothed base model prediction, and 3 the improved patch-based model prediction. Image 4 shows the base model absolute T2 Z-score (base model best performance) and a Z-Score overlay with minimum threshold set to 4 and maximum threshold set to 10 and image 5 the 7x7x7 surface patch model of the averaged (mean) T1w and T2w absolute Z-scores (best overall performance) also with minimum threshold set to 4 and maximum threshold set to 10.

Combining the patch-based models and iterative within image lesion infilling helped maintain lesion sensitivity and preserved inclusion of subject-specific anatomy. Due to this, patch-based model outputs were more interpretable as they produced synthesised images that were anatomically closer to the subjects observed MRI image, allowing better association between specific areas of the brain, Fig 4.3 1-3 (Rutherford et al., 2010; Volpe, 2012). The only areas this did not hold true were for lesioned areas, which reverted to the prior due to being unable to estimate the true appearance of the underlying healthy tissue structure from the neonates actual MRI. However, given that the underlying normal structure cannot be inferred due to the damage present in the subjects MRI, these more general interpretations are acceptable. In their work using adversarial auto-encoders, Chen & Konukoglu, 2018 produced synthesised images that were both blurrier and included more pathologic tissue as did similar work by Zimmerer et al., 2019. In their review of adversarial auto-encoder based methods Baur et al., 2021 found this was an issue shared by many methods attempting creation of synthetic images. Frequently,

a trade-off between inclusion of pathological tissue, image sharpness and correct inclusion of heterogeneous normal structure had to be made.

Table 4.7 Neonates from the second brain injury cohort that were correctly identified by the different 7x7x7 surface patch models. Rows 1-3 show models created from the T1w Z-score histograms, T2w Z-score histograms and mean of the T1w and T2w Z-score histograms. The 4th row shows the mean of the probabilistic output from the separate T1w and T2w Z-score histogram models. Each cell in the Abnormal and Normal Image columns shows the number of correctly identified images followed by the total number of images of that type. Values in bold indicate the best overall performance.

Model	Abnormal Images	Normal Images	Sensitivity	Specificity
T1w Z-score	56/73 (76.71%)	62/73 (84.93%)	0.767	0.849
T2w Z-score	60/73 (82.19%)	61/73 (83.56%)	0.822	0.835
Mean T1w and T2w Z-scores	54/73 (73.97%)	50/73 (68.49%)	0.740	0.685
T1w Prediction + T2w Prediction	64/73 (87.67%)	62/73 (84.94%)	0.877	0.849

Of the selection of patches tested, the 7x7x7 surface patch iterated three times was the best performing model. It produced accurate predictions and attained a higher AUC (0.961) and, for a fixed specificity, had higher sensitivity scores for most thresholds. While the 7x7x7 dense patch model did produce a higher sensitivity score at the largest threshold, it is likely that this is due to the effect of rounding the scores. These sensitivity scores were in-line with those seen in the deep learning-based methods used by Essa et al., 2020 but lower than those seen in N. Subbanna et al., 2015, although neither presented specificity scores for comparison. Roy et al.,

2018 did not publish sensitivity scores, though they did publish dice coefficient, which were also higher. A point to remember is that, in all three cases, these were applied to a single type of pathology. Methods covered in the introduction chapter, (Srhoj-Egekher et al., 2012; Guizard et al., 2015; Steenwijk et al., 2013; Kniep et al., 2019; Subudhi et al., 2020) achieved sensitivity scores in the range of 0.601 to 0.925 but, were all applied to specific pathologies and provided no specificity scores.

The review by Baur et al., 2021 compared 18 different adversarial autoencoders, all of which base their detection on outliers from expected normal images. These were applied to four different pathological datasets: one glioma and three multiple sclerosis and AUC scores were available for all. Only one method, Baur et al., 2019, achieved AUC scores greater than my own method, with another, (Zimmerer et al., 2019), doing so once in one dataset. This demonstrates performance on par with state of the art performing deep learning methods.

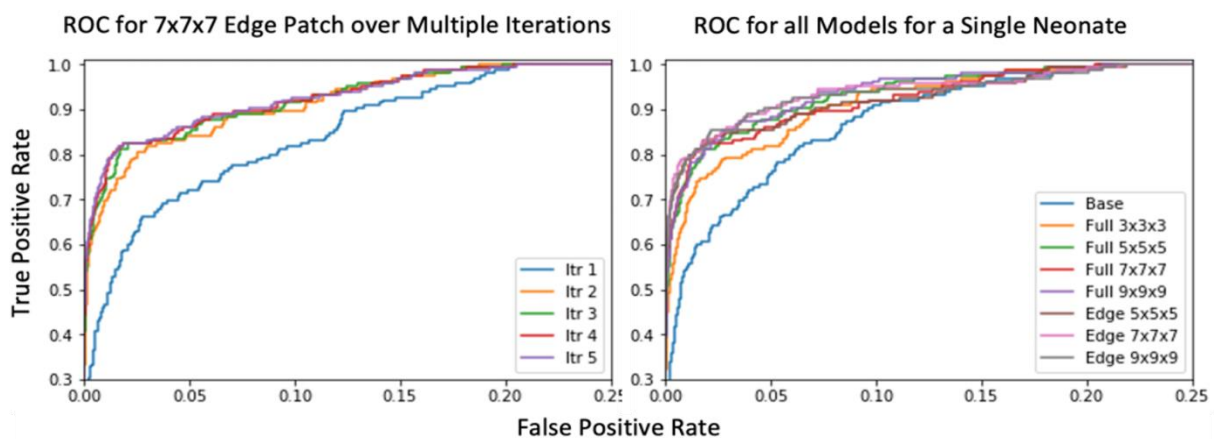


Figure 4.4 Improved ROC performance in patch-based models. Left image: ROCs for the same neonate for multiple iterations of the 7x7x7 Edge Patch model, right image: Preterm ROCs for all models for a single neonate.

In general model performance was enhanced both via inclusion of patch information and by iteration of the model. This can be seen in Fig 4.4 which gives an example of a single neonate with germinal matrix haemorrhage whose AUC improved with use of a patch-based model and over multiple iterations. The absolute Z-score overlays from only the T2w and mean of the

T1w and T2w 7x7x7 surface patch models, both with a threshold of 4 to 10 applied, are shown in Fig 4.3 4 and 5 respectively. This demonstrates an improvement in the segmentation coverage of the germinal matrix haemorrhage using the mean of the T1w and T2w absolute Z-scores.

Using the mean of the T1w and T2w absolute Z-score maps increased the AUC for all patch models compared to single modality absolute Z-score models. The higher AUC scores are likely due to an increase in specificity via reduction in the number of noisy (false positive) voxels since it is unlikely that the same random voxel will be noisy in two images. While good for noise reduction and for pathologies with a strong presence in both modalities, it opens the possibility that certain pathologies, which may have a more subtle appearance in one modality, having poorer lesion coverage (Barkovich, 1998; J. H. Miller et al., 2020; M. Rutherford, Biarge, et al., 2010).

Patch-based models were better at detecting both smaller pathologies and those with more subtle intensity differences. Small haemorrhages or groups of PWML had more complete coverage of lesioned tissue at the same absolute Z-score threshold when compared to the base model. Some pathologies, e.g., a single PWML, tiny cerebellar haemorrhage (2-4 voxels), could not be easily detected. However, these were almost indistinguishable due to partial volume or blurring during registration which altered their intensity sufficiently to blend in with normal tissue. In many instances these could only be easily identified by close examination of the raw images.

However, patch models had lower performance when looking at some larger pathologies due to the inclusion of abnormal tissue into the predicted image. This is attributable to two factors. First is that the model will partially infer the intensity of the target from the voxels included in the patch as, in most cases, target voxels contain intensities similar to those surrounding them.

As such, if pathological tissue intensities are included in the patch, the model will also attempt to infer from these, similar to that seen in methods described in (Baur et al., 2021). Second, if there is a training image patch with intensities close to those of the abnormal tissue in the target patch it will interpolate from that image. Due to this, large lesions (like cystic periventricular leukomalacia) often miss the lesion core but detect the edges.

An argument can be made here that this is an acceptable exchange for better coverage of smaller lesions. Missed lesions, even small ones, could have a significant effect on the immediate and long-term outcome of the neonate depending on their location (Prempunpong et al., 2017; Rollins et al., 2014). Early detection of these injuries, that could otherwise be missed, could help alleviate these negative effects or find residual injury traces after intervention. Even though the core of larger lesions is missed, detection of lesion edges is still sufficient to bring them to the attention of radiological staff reviewing the images. As the goal of this system is to aid staff in detection and assessment of injury, not an absolute alternative, better general coverage of all lesions is a more important result. Final judgement about the extent and consequence of the injury can be made by the clinical staff once they have been made aware of its presence.

The surface patch model has several advantages over the dense patch model. Excluding the voxels closest to the target voxel reduces the chance that pathological intensity values will be included in the design matrix. This makes it less prone to including lesioned tissue in the predicted image. Despite this loss of close structural information, it is still able to infer tissue structure from surface voxels which, depending on the size of the patch, are only two, three or four voxels away from the prediction target in all directions. It also has the advantage of computational efficiency by including only the outer most voxel intensities into the design matrix. This both reduces the dimensionality of the input space, requiring less training sample to adequately sample, and the complexity of the hyperparameter optimisation.

When applied to the second held out brain injury dataset, model performance remained consistent with that seen in the initial preterm cohort. The only major difference was that at an absolute Z-score threshold the 5x5x5 surface patch had slightly better performance than the 7x7x7 surface patch. That said the performance between the two models was very similar.

For the classification task, making predictions using separate T1w and T2w Z-score models and combining their probability prediction outputs achieved the best results. This identified 32/36 abnormal and 28/33 normal images, outperforming both individual modality classifiers. Sensitivity scores in Kaplan et al., 2022 were higher, $\approx 99\%$, but had an unbalanced dataset favouring pathology, $\approx 5/1$, and classified only Parkinson's disease. Classification accuracy in the methods reviewed in the introduction chapter, (Zurita et al., 2018; Saccà et al., 2019; Pinto et al., 2020; Rallabandi et al., 2020; Richhariya et al., 2020; Velazquez et al., 2021; Vamvakas et al., 2019; Kaur et al., 2019), ranged from 0.73 to 1 and specificity scores ranging from 0.667 to 0.96. This again demonstrates my model's ability to match many state-of-the art techniques, even given its application to multiple types of pathology.

Some images were missed, but these contained single/very small PWML and small cerebellar haemorrhages with subtle intensity differences as noted in chapter 3. The results from the second brain injury cohort once again supported those found in the initial preterm cohort. Combining the probabilistic output of the separate T1w and T2w classifiers achieved the best result, correctly identifying 64/73 abnormal images and 62/73 normal images. This supports the argument that pathologies are more or less apparent on different modalities (M. Rutherford, Biarge, et al., 2010; Tusor et al., 2017) and that using some combination from multiple acquisitions is beneficial. Interestingly, the classifier using the histogram from the mean of the T1w and T2w Z-scores gave the worst performance, being poor at identifying neonates with normal appearing brains.

The primary issue affecting the patch-based models was the inclusion of lesioned tissue in the predicted image, a situation shared by other methods that used synthesized normal images (Baur et al., 2021). This could be corrected by better standardization of the tissue intensities of individual images. This would help prevent the fringe case issues where, in some normal appearing neonates, normal tissue has intensities closer to those found in pathological tissue in other neonates. Alternatively, average intensity values could be created from tissue segmentations and incorporated into the model, used to group similar images together or to correct initial segmentation via a second method that removes false positives (P. K. Roy et al., 2015; Saladi & Amutha Prabha, 2018; Srhoj-Egekher et al., 2012.; N. Subbanna et al., 2015; N. K. Subbanna et al., 2019). A final and potentially more complex solution would be some method of dynamically removing or altering the contribution to the prediction model of voxels within a patch with abnormal intensities.

The logistic regression classifier used for this work represents only the initial steps into this area and only acts as a proof of concept with multiple methods of improving its efficacy. These include using a more advanced non-linear or non-parametric method that can better separate the input data such as a random forest, support vector machine, GP classifier (Salvador et al., 2017) or potentially some form of deep learning architecture. The task itself could also be refined. This could include optimising for a more specific type of pathology, for instance ensuring 100% sensitivity of any haemorrhage or stroke cases. It could also include expanding it to a multi-output framework, such as not only identifying pathological cases but estimating their severity.

A final note is that both the GP model and the classifier could be improved with access to larger, more comprehensive datasets. For the lesion free neonates this would allow for a more accurate estimate of typical development. With lesioned neonates, it could give more robust

estimation of lesion detection and coverage as the current small sample size leaves it susceptible to under-estimation errors (Varoquaux, 2018).

4.5 Conclusion

The inclusion of patch information into MRI normative models results in synthesized “typical” images that more closely match subject specific structural heterogeneity while maintaining excellent lesion detection. Optimal performance was attained by a robust examination of different patch types and sizes combined by an iterative application of the model. Of those tested, the 7x7x7 surface patch with prediction iteration improved abnormality detection giving higher AUCs, a better sensitivity specificity trade-off and better coverage of smaller lesions. The mean of the probabilistic output from the separate T1w and T2w absolute Z-score histogram models was able to correctly identify 32/36 abnormal images in an initial preterm cohort and 64/73 abnormal images in a separate brain injury cohort. This could act as a screening tool, only missing those images with very small lesions that were almost imperceptible after non-linear registration.

Chapter 5. Automatic Detection of Hypoxic Ischemic Encephalopathy using Outlier Detection

Hypoxic ischemic encephalopathy (HIE) affects 1.5 in every 1000 term neonates, where those with more severe injury patterns frequently develop negative long-term functional outcomes such as poor cognitive ability and cerebral palsy. HIE has known injury patterns, whose changing appearance in T1 weighted (T1w), T2 weighted (T2w) and apparent diffusion coefficient (ADC) images over time is well documented. This chapter proposes the use of the methods developed in chapters 3 and 4 of this thesis for within image detection of lesions associated with HIE and across image classification of neonates with HIE. It also investigates the use of these methods for the more challenging task of prognostic classification, specifically predicting if a neonate with a confirmed diagnosis of HIE had a typical or poor long-term functional outcome. Finally, I examined different enhancements to the classification. This included the use of additional classification algorithms that see common use in neuroimaging challenges. Specifically, these were a random forest, two support vector machines (linear kernel, radial basis function kernel) and a Gaussian process classifier. A local estimation of lesion abnormality, the regional abnormal voxel count (RAVC), which counts the number of abnormal voxels over a given thresholds, was used as input to these classifiers alongside the previous method based on histograms of the Z-score maps. The datasets for HIE vs healthy control (HC) and HIE normal vs poor outcome classifiers were composed of 62 neonates with HIE and 52 term born healthy controls (HIE vs HC), and a subset of 17 HIE neonates with poor outcomes and 15 HIE neonates with good outcomes (HIE normal vs poor). The best performing model for the HIE vs HC model was a linear SVM using RAVC scores from ADC images which correctly identified 58/62 neonates and 48/52 healthy controls (Sensitivity 0.935, specificity 0.923). For the HIE normal vs poor outcome classifier, the RF model using RAVC scores from T2w images gave the best performance, correctly identifying 14/17 neonates with

poor outcomes and 11/15 neonates with typical outcomes (sensitivity of 0.824, specificity 0.733). These results demonstrate the potential of simple, two-stage analysis of clinical imaging data for predicting outcome in clinically challenging HIE.

5.1 Introduction

Hypoxic Ischemic Encephalopathy (HIE) is a condition affecting term born neonates whose brain is starved of oxygen and/or blood just prior to, during or just after birth. The severity and extent of injury observed in HIE is correlated with both the duration and acuteness of the hypoxic ischemic insult. Longer duration and more complete interruption result in greater levels of injury (Krishnan & Shroff, 2016). Incidence rates of HIE are 1.5 neonates per 1000 term births in the UK (Kurinczuk et al., 2010). Clinical intervention via hypothermic cooling has been shown to mitigate the occurrence and extent of negative outcomes (Shankaran et al., 2005). This makes early identification of HIE cases within newborns critical to assuage negative effects and improve long-term outcomes.

An initial HIE diagnosis is dependent on a combination of several known symptoms. These can include low umbilical cord blood gases, need for resuscitation, systemic organ failure, seizures, abnormal heart rate, respiratory failure, reduced consciousness, reduced movement, tone and posture abnormalities, feeding difficulties and general distress. Additional clinical indicators included umbilical cord acidosis scores of <7 or <7.1 (depending on institution) and Apgar scores of below 10 at 2, 5 and 10 minutes (Bano et al., 2017; Gopagondanahalli et al., 2016; Guo et al., 2016; Krishnan & Shroff, 2016; Millar et al., 2017; Piñeiro-Ramos et al., 2020; M. Rutherford, Malamateniou, et al., 2010; Schiffrin & Ater, 2006; Volpe, 2012). Alternatively, HIE can present after a specific sign, known as a sentinel event, such as bradycardia, umbilical cord knotting or pneumonia (Fineschi et al., 2017; Krishnan & Shroff, 2016; Shankaran et al., 2017; Volpe, 2012).

While these symptoms and/or events may give an indication that some form of HIE injury may have occurred, it is not yet a concrete diagnosis. Alongside ongoing monitoring of the neonate's clinical condition, there is frequently follow-up imaging. This allows in vivo observation of the neonatal brain to confirm the presence of injury via some form of associated and identifiable imaging characteristic. One increasingly used method of confirming and evaluating an initial HIE suspicion is the use of Magnetic Resonance Imaging (MRI) (Rutherford et al., 2010).

Another important issue associated with HIE is its effect on a neonate's long-term outcome. There are several well documented conditions with associations to neonatal HIE. These include cerebral palsy, motor, developmental, speech and cerebral delay or impairment and in the most extreme cases death (Perez et al., 2013; Perlman & Shah, 2011). HIE is often split into mild, moderate and severe categories depending on the location and extent of the tissue damage and more severe injury is often associated with poorer outcome.

Within the spectrum of possible injuries associated with HIE, several patterns show greater prevalence. These include damage to the basal ganglia and thalamus (BGT), cortical grey matter, subcortical white matter and, in more severe cases, the brain stem and cerebellum (Gopagondanahalli et al., 2016; Millar et al., 2017; Quattrocchi et al., 2016; M. Rutherford, Malamateniou, et al., 2010; Volpe, 2012)

White matter damage may include diffuse white matter injury, where large areas of the cortical white matter suffer damage to the oligodendrocytes. Small focal lesions, similar to punctate white matter lesions (PWML), have also been observed; frequently in lines adjacent to the lateral ventricles (A. M. Li et al., 2009). Larger focal white matter lesions located around the ventricles also become more frequent and severe. In severe instances oedema can occur causing cell death and leading to periventricular leukomalacia (PVL) (Volpe & Zipurksy, 2001).

BGT lesions can include haemorrhage, oedema, myelin destruction around the posterior limb of the internal capsule and eventual atrophy of neural tissue. Damaged cortical grey matter areas can also undergo oedema and atrophy resulting in large extracortical space and gyri having a ragged appearance (Rutherford et al., 2010). Damage to the brainstem often appear as column shaped lesions running throughout its structure (Quattrocchi et al., 2016). In extreme cases, complete cerebral devastation can occur, wherein almost all of the cortical grey matter and subcortical white matter atrophy in combination with extensive lesions in the BGT (Shankaran et al., 2012).

There are two additional factors that must be considered when assessing HIE injury, the modality in which the image was acquired and the timespan between suspected injury and the scan (Rutherford et al., 2010; Volpe, 2012). T1 weighted (T1w), T2 weighted (T2w) and some form of diffusion image are frequently acquired in clinical neonatal imaging. In this context, the diffusion image is usually displayed as an apparent diffusion coefficient (ADC), also known as a mean diffusivity, image.

When the BGT is abnormal, the appearance varies depending on severity and time after insult. Initial damage has the same appearance as the cortical surface, hypo intense in T1 and hyperintense in T2. However, due to greater reliance on oxygen, damage to deep grey matter progresses more rapidly (Imai et al., 2018; Pasternak et al., 1991). This can result in haemorrhages, with contrast dependant appearance, alongside oedema and cell death, resulting in CSF filled areas of damage. These injuries often result in patterns described as having a “marbled” appearance. Similar to other modalities, ADC positive injuries rely on context dependant changes in intensity, although in this instance this represents different diffusivity levels than would be observed in a healthy neonate. Any fluctuation in the level of diffusivity, regardless of whether it is an increase or decrease, can be relevant and may be different depending on the time after injury (Wolf et al., 2001).

The acute phase of HIE is described as occurring over a 2-week window following birth/injury with imaging recommended during the second week (Rutherford et al., 2010). Waiting for up to two weeks for this assessment is not an ideal course. This is where ADC images become incredibly useful, as lesions, if present, become visible much sooner (Wolf et al., 2001). Using ADC images, HIE can be detected as early as a day after the initial injury as areas of lower-than-expected diffusion. This quick detection is crucial for fast confirmation of the HIE diagnosis and early intervention.

It should be noted that while useful for early detection purposes, there are challenges with relying purely on ADC based diagnosis. Over time, abnormal signals present in ADC images can undergo a process known as pseudo-normalisation, wherein areas of abnormal intensity return to their typical values. This process progresses at different rates depending on the type of injury, its location and severity. As a result, there can be underestimation of the total lesion area depending on precisely when the ADC image was acquired. This pseudo-normalisation process begins around day 7 and is completed by day 10-14 with diffusion intensity values potentially completely returning to pre-injury levels (Bano et al., 2017; Krishnan & Shroff, 2016; Rutherford et al., 2010).

The use of T1w and T2w acquisitions complement these ADC image findings over the latter part of this time period. Although an injury may start to become visible in T1w and T2w images within 3-7 days, it is after this, when pseudo-normalisation occurs in ADC, that lesions become most apparent, allowing for more accurate assessment of total injury (Cowan et al., 2003; Millar et al., 2017). This is due to the injury having a different time associated appearance depending on its stage; the initial HIE injury, a recovery phase (dependant on severity of initial injury) and finally, if severe, cell death (secondary energy failure) (Gopagondanahalli et al., 2016). To summarise the injury progress in T1w/T2w; from day 3-7 lesions appear as areas of

swelling and/or loss of MRI contrast/detail, and by day 7-10 injury type and extent should be clear and more readily assessable.

Damage to the BGT is often highlighted as being associated with severe motor and cognitive impairment and cerebral palsy even in milder cases, with more severe injury resulting in greater impairment (Rutherford et al., 2010). A retrospective study by Martinez-Biarge et al., 2011 found that of 361 neonates with some form of HIE, 175 contained lesions in the BGT, of these 49 died and 89 went on to develop cerebral palsy. The appearance of the posterior limb of the internal capsule, which separates the thalamus and globus pallidus within the BGT, is also noted as being important for diagnosis, with higher levels of degradation associated with poorer outcomes. Cortical surface injury can result in delays or impairment to their associated function, although to a lesser extent compared to injuries in the BGT. Damage to the subcortical WM is, by itself, usually a milder injury and associated with better outcomes but will exacerbate long term negative effects if seen in conjunction with BGT and cortical lesions. Brainstem injury is seen in only the most severe cases and many neonates with this level of injury do not survive. Prediction of long-term effect is difficult, as while some injuries are known to be associated with more severe effects, such as damage to the BGT, this is not always the case. In a similar manner, even neonates with relatively normal appearing brains, with little to no lesion content, will go on to develop some form of delay or impairment.

Automatic detection of HIE is a relatively new field with only a small number of publications available, and which are often reliant on non-imaging information. Of those that do use MRI, Tang et al., 2020 applied ridge detection to 42 susceptibility weighted images of neonates with HIE to classify outcome severity (31 normal, 11 poor outcome). Murphy et al., 2017 used diffusion weighted images (ADC) from 20 HIE neonates as inputs to a random forest classifier (RF) for voxel-wise classification and achieved high sensitivity and specificity values of 0.72 and 0.99 respectively. Weiss et al., 2019 laid out a study program to identify potential

biomarkers relevant to HIE detection and outcome prediction to a multi-site cohort of over 300 neonates but have not, at time of writing, published results.

Several methods use electroencephalography (EEG) to detect, Dereymaeker et al., 2019 (19 neonates) and Stevenson et al., 2013 (54 neonates), or quantify the severity and Ahmed et al., 2016 (54 neonates), of HIE. The issue with purely EEG based methods, is that while they can detect gross disruption of typical EEG signal, they cannot localise this damage to specific areas of the brain, an area in which MRI could assist.

More recently there has been increasing interest in normative models, Finck et al., 2021; Macleod et al., 2021; O’Muircheartaigh et al., 2020; Weiss et al., 2019 and Ziegler et al., 2014, that use deviation from a typical healthy reference group, otherwise called outlier detection, to highlight abnormality. These methods have seen further use for these abnormality maps as inputs for either separation of healthy controls (HC) and pathological cases, as was seen in chapter two of this work, or as outcome prediction after a specific pathological insult, Weiss et al., 2019.

Here, I propose the use of a multi-step detection method to identify HIE in the neonatal brain, initialised from my work in chapters 3 and 4. Initially I will use Z-score maps, generated from neonatal MRI alongside voxel-wise GP synthetic images and standard deviation maps, as input for diagnosis classification (HC vs HIE) based on anatomically informed priors. I will then estimate the utility of this method for the more challenging task of prognosis; identification of neonates with HIE that had negative outcomes, similar to that suggested in Weiss et al., 2019. I will also compare the efficacy of several commonly used machine learning techniques for each of these classification tasks.

5.2 Methods

5.2.1 Sample and dataset

A neonatal cohort comprising 254 individuals was acquired from a study approved by the West London Research Ethics Committee, and written consent was obtained from the parents of all the infants before scanning. These included T1w, T2w and ADC images acquired on a Phillips 3T scanner, the acquisition parameters for which are shown in Table 5.1. This included both healthy controls (HC, n=108), and neonates with HIE (n=146), with gestational and postmenstrual ages (GA, PMA) between 37.5 and 50-weeks.

Table 5.1 Acquisition parameters.

Scan	TR	TE	Voxel Size
T2w	7000-16000ms	160ms	0.859 x 0.859 x 1mm
T1w	17ms	4.6ms	0.8 x 0.8 x 1.6mm
T1w High Resolution	17ms	4.6ms	0.8 x 0.82 x 0.82mm
32-Direction Diffusion	8498.1ms	49.ms	1.75 x 1.75 x 2 mm

A comprehensive manual review of all images was undertaken to assess image quality. Any images with the following artifacts were removed, inconsistent slice intensities, strong inhomogeneity artifacts or motion blur. I also ensured that each subject had a T1w, T2w and Diffusion image and that there was no missing demographic data, which substantially reduced sample size. For this work, only diffusion datasets acquired in 32 direction and with an accompanying b0 image were used, neonates with diffusion acquisition in less directions (usually 16) or no b0 image were excluded. This resulted in a final sample of 114 neonates, where 52 were part of the HC group used for normative modelling of the term neonatal brain

and the remaining 62 were pathological HIE neonates with varying levels of injury, summarized in Table 5.2.

Table 5.2 Cohort information.

Parameter	Value
Total Subjects (Female)	114 (59)
Post Menstrual Age (PMA) Range (Weeks+Days) [Median]	37+2 to 50+0 [41+6]
Gestational Age (GA) Range [Median]	36+0 to 41+6 [39+6]
Healthy Controls (Training)	52(25)
PMA Range [Median]	38+2 to 50+0 [43+0]
GA Range [Median]	36+0 to 41+6 [39+0]
Hypoxic Ischemic Encephalopathy Neonates	62(34)
PMA Range [Median]	37+2 to 46+4 [41+0]
GA Range [Median]	36+0 to 41+6 [40+2]

Within the HIE neonates, 32 had some form of follow-up to assess subject outcome, of which 15 had typical outcomes and 17 had a negative outcome, shown in Table 5.3. These negative outcomes were death (2), cerebral palsy (8), or some form of delay or impairment (developmental 15, speech 3, motor 4, behavioural 1 and/or cognitive 2). Some neonates had multiple categories of negative outcomes. Conversely, those neonates defined as being typical had no form of delay, impairment or morbidity at time of follow-up.

Table 5.3 HIE cohort with outcome information.

Parameter	Value
HIE Neonates with Outcomes (Female)	32 (19)
Post Menstrual Age (PMA) Range (Weeks+Days) [Median]	37+2 to 46+4 [41+3]
Gestational Age (GA) Range [Median]	36+0 to 41+6 [40+2]
HIE Neonates with Normal Outcome	15(9)
PMA Range [Median]	37+3 to 43+2 [41+5]
GA Range [Median]	36+6 to 41+6 [41+0]
HIE Neonates with Poor Outcomes	17(10)
PMA Range [Median]	37+2 to 46+4 [41+2]
GA Range [Median]	36+0 to 41+6 [40+2]

5.2.2 Pre-processing

FSL (Jenkinson et al., 2002; Jenkinson & Smith, 2001) was used to rigidly register neonatal T1w images to T2w images followed by B0 inhomogeneity correction with N4BiasFieldCorrection (Avants et al., 2014) and removal of extracranial tissue using FSLBET (Smith, 2002). FSL was also used to create affine registrations between T2w images and weekly templates (38-44 weeks) from the developing human connectome project (dHCP) (Makropoulos et al., 2018; Schuh et al., 2018). These were used as initial transforms for an ANTS (Avants et al., 2014) based non-linear registration to the same templates. These

individual affine matrices and warps were combined with non-linear warps between individual week templates and the week 40 template. These were then applied to both the T1w and T2w images to ensure they occupied the same space. Voxel intensities within each image were standardized to image intensity median and standard deviation as explained in chapter 3 of this thesis.

The 32 direction diffusion images were first visually reviewed and any image with more than 8 corrupted volumes discarded. All remaining images had corrupt volumes removed. The ADC images were constructed using FSL BET, eddy current correction (Andersson & Sotiropoulos, 2016) and dtifit. ADC images then followed the same alignment to template space as the T1w images, first a rigid registration to their T2w image counterpart followed by application of the combined affine matrix and nonlinear warps.

Subject demographics were transformed in order to assist the modelling process. First the sex was binarized; male as 0 and female as 1. The GA and PMA both underwent standardization using the median and standard deviation of the combined HIE and HC dataset.

5.2.3 Generating Subject Specific Predicted and Outlier Z-score Maps

As described in chapters one and two, Gaussian Process (GPs) models were fit at every voxel location within a brain mask, where each GP will output the intensity mean and variance for that voxel. Each GP uses the GA, PMA and sex as well as a patch of voxels surrounding the prediction target as inputs. This patch of voxels is used to infer anatomical context from the surrounding tissue in order to help accommodate individual structural heterogeneity. This is based on the optimal patch seen in chapter 4 (Macleod et al., 2021), where the outermost voxel intensities from a 7x7x7 patch are appended alongside the GA, PMA and sex. A more detailed description can be seen in chapter 4 section 3.2. Modelling was achieved using GPyTorch (Gardner et al., 2018) applied to the HC dataset.

For truly comprehensive modelling over this age range and to account for brain heterogeneity, a larger cohort would be necessary. However, this is offset by use of the patch-based model, which incorporates increased spatial information and allows for direct inference of voxel intensity values.

The optimisation (training) for the GP model hyperparameters is performed on the term born healthy control dataset. After optimisation, these images also provide a normative dataset of input parameters and corresponding normal intensity values for each voxel. These are used by the GP models to determine the similarity of the target input parameters to the sample (HC training) parameters for inference of the output variable.

By combining the mean outputs for every GP voxel prediction within a brain mask, a predicted image for a given neonate of a specific age, sex, and image contrast can be reconstructed alongside a corresponding variance map. By iterating the model using patches from these predicted images, pathological tissue intensities (outliers) get progressively pushed back to this expected mean of the training data, effectively “infilling” pathological images. This compensates for the high variance values that occur in voxels with abnormal intensity values, as shown in chapter two.

With access to these newly synthesised normal appearing brain images, the corresponding variance maps and the original neonatal images, individual Z-score maps can be created which have superior sensitivity and specificity. This is achieved by finding the difference between the subjects original and predicted image and then scaling it by standard deviation maps converted from the original variance maps. This gives a voxel-wise measure of abnormality, as performed in previous chapters, where values which are increasingly far from zero had MRI voxel intensities further than expected from typical voxel values and thus more likely to be abnormal.

5.2.4 Injury Localization

The dHCP templates included weekly regional atlases which are aligned to their template counterparts (Schuh et al., 2018). As all subject images are registered to the week 40 template, the subject biological regions show a good overlap with those on that atlas. These regions can be overlaid over subject Z-score maps in order to find how many voxels within a specific region are over a threshold. As such, a new measure called regional abnormal voxel count (RAVC) is created for use as inputs to the classification models. A total of 85 regions are included in the atlas shown in Table 5.4, a multiplane cross section of the atlas is shown in Fig 5.1.

Table 5.4 The regions present in the dHCP based atlas and their different subdivisions.

Main Structure	First Sub-division	Second Sub-division	Third Sub-division
Frontal lobe	Left, Right	GM, WM	
Parietal lobe	Left, Right	GM, WM	
Occipital lobe	Left, Right	GM, WM	
Anterior temporal lobe	Left, Right	GM, WM	Medial, Lateral
Superior temporal gyrus	Left, Right	GM, WM	Middle, Posterior
Medial+Inferior temporal gyri	Left, Right	GM, WM	Anterior, Posterior
Lateral occipitotemporal gyrus, Gyrus fusiformis	Left, Right	GM, WM	Anterior, Posterior
Insula	Left, Right	GM, WM	
Cingulate gyrus	Left, Right	GM, WM	Anterior, Posterior

Corpus callosum	Left, Right		
Gyri parahippocampalis et ambiens	Left, Right	GM, WM	Anterior, Posterior
Hippocampus	Left, Right		
Amygdala	Left, Right		
Thalamus	Left, Right		High, Low intensity
Subthalamic nucleus	Left, Right		
Caudate nucleus	Left, Right		
Lentiform nucleus	Left, Right		
Lateral ventricle	Left, Right		
Cerebellum	Left, Right		
Brainstem			

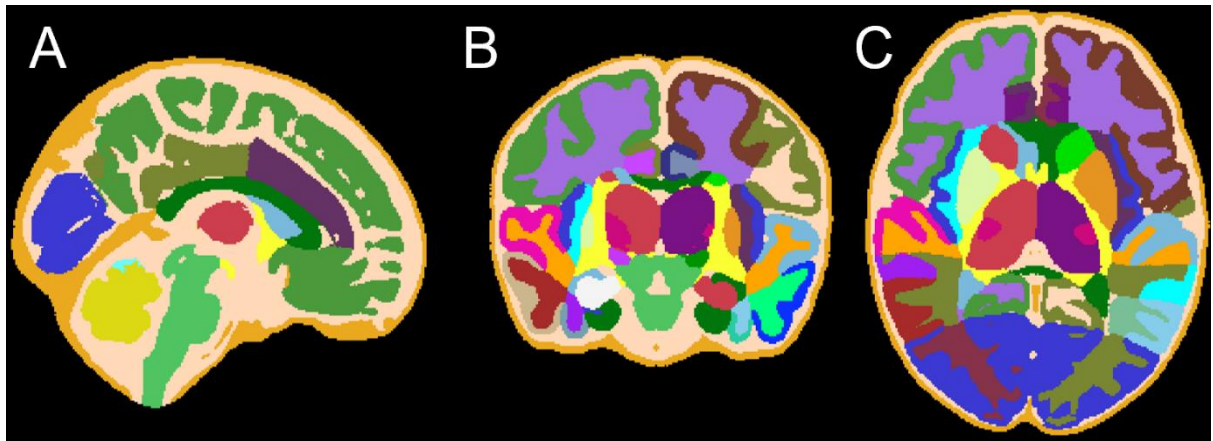


Figure 5.1 Images of the 85-region atlas used to create the regional abnormal voxel count. A threshold is applied to a Z-score map and the number of voxels in each region over that threshold are counted. Column A shows the sagittal plane, B the coronal plane and C the axial plane.

5.2.6 Classification Algorithms

A more detailed explanation of the implementation and operation of the following classification techniques is given in introduction chapter 2 of this thesis and the default parameter are show in Table 5.5.

Table 5.5 Classification methods used in this chapter and their parameters.

Model	Default Parameters
LR	Penalty = L2 Tolerance (stopping criteria) = 0.0001 C (regularisation) = 1 Intercept = True Solver = limited memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) Max (solver) iterations = 1000 Multiple Class = Multinomial

LIN-SVM	C = 1 Tolerance (stopping criteria) = 0.0001 Cache (kernel memory) size = 200MB Max iterations = 1000000 Multiple Class = One vs All Kernel = linear
RBF-SVM	C = 1 Tolerance (stopping criteria) = 0.0001 Cache (kernel memory) size = 200MB Max iterations = 1000000 Multiple Class = One vs All Kernel = radial basis function (RBF) Gamma = scale
RF	Number of estimators (Trees) = 100 Criterion (split quality estimator) = Squared error Max depth = None Minimum samples for split = 2 Minimum leaf samples = 1 Max features (for split) = Square root (of total samples) Max leaf nodes = None Min impurity decrease = 0 (will split if impurity decreases) Bootstrap samples = True Cost-complexity pruning = 0 (no pruning of trees)
GPC	Kernel = RBF

	Optimiser = LBFGS Optimiser restarts = 100 Predicter iterations = 100 Copy inputs = True (keeps input values stored in memory) Multiple Class = One vs All
--	---

Briefly, LR models operate by applying a weighting to each covariate which are then added together giving a linear combination of these inputs. These are then passed through a logistic (sigmoid) function and outputs over 0.5 are labelled as 1 and those below as 0. The optimisation process changes the values of these weightings so that the results of the linear combination allow different classes to be separated around this 0.5 value, in this instance using L-BFGS optimiser (Zhu et al., 1997). In the formula 5.1, y is a row vector of target classes (0,1), X is a matrix of inputs where each row is a sample and each column a covariate and θ is a vector of weights.

$$y = \frac{1}{1 + e^{-(x^T \theta)}}$$

5.1

SVMs are wide margin classifiers that attempt to separate two (or more) groups of data by applying a boundary line between them. The SVM then maximises the space between this line and the nearest samples from each class in the training data (the support vectors) based on the type of kernel used. In the linear case, formula 5.2 where i is a single datapoint and n is the total number of datapoints, this a hard margin that fits a line that finds the best split within the data. For the RBF-SVM it applies the soft-margin seen in formula 5.3. This is combined with a projection into higher-dimensional space using the kernel trick to allow better separation of the data.

$$y_i(\theta^T x_i - b), \text{ for all } 1 \leq i \leq n$$

5.223

$$\max(0, 1 - y_i(\theta^T x_i - b))$$

5.3

RF classifiers are composed of multiple decision trees. Each tree selects different subsets of samples and covariates from the training set and will attempt to separate the output classes by repeatedly dividing the samples based on some covariate. The overall classification of a sample is given by the combined output of all trees within the forest. The Gini impurity (Quinlan, 1986) is used to estimate the effectiveness of every split within all trees. Each decision tree, formula 5.4, has a maximised split weighting W for each feature f for each parent node n , where X_n are the parent node samples, X_l and X_r are the left and right child node sample subsets, and I is the Gini impurity.

$$W(X_n, f_i) = I(X_n) - \frac{X_l}{X_n} I(X_l) - \frac{X_r}{X_n} I(X_r)$$

5.4

The gaussian process classifier (GPC) operation is the same as specified for GP regression, but in this instance applies a logistic function to the output. This changes the output from continuous values into discrete sets while still preserving the probabilistic output associated with GPs.

Several different combinations of variables were used as inputs to these classifiers. The initial group of models used histograms from voxel-wise Z-scores of the T1w, T2w and ADC images as well as a fourth created by combining the T1w and T2w histograms. The histograms contained 21 equally spaced bins from -20 and 20 with a spacing of 2.

An additional set of models were built by overlying the multi region atlas onto each subject T1w, T2w and ADC images and the RAVC for each region over a given threshold found. However, to begin with, the correct threshold for each image acquisition type's Z-score is unknown and must be optimised. To achieve this, 100 models were built for each classifier, modality with increasing thresholds applied to the Z-scores, an effect equivalent to calculating the ROC. As the threshold increases, the RAVC score decreases as less voxels meet this required Z-score threshold, removing those that are more likely to be normal first. In abnormal images there will be more voxels with higher Z-scores and therefore a higher RAVC. These thresholds start at 0.5 and increase in increments of 0.5 up to a maximum threshold of 50. Each threshold model had its accuracy assessed, allowing for the identification of the best performing threshold by comparison. After finding the best threshold for the T1w and T2w Z-scores, these were combined into a further model for each classifier as was done with the histogram-based models. It was decided against combining information from the ADC images as while the injury patterns in T1w and T2w images can be similar, albeit with differing intensities, the ADC images can vary greatly due to the changing injury pattern over time. The averaging effect over the three modalities would most likely prevent detection unless the lesion had a strong presence in all three modalities. The best performing model for each classifier and acquisition modality can then be compared to the equivalent histogram-based models.

5.2.7 Estimating Model Performance

In order to gauge each model's performance, the sensitivity and specificity of each model was calculated in a 5-fold cross validation set. Sensitivity is the percentage of cases which have been correctly identified, in the first task those diagnosed with HIE and the second as those having negative outcome and is represented as values of 1 (positive). specificity is the percentage of cases that have been correctly identified as being either HC, for the first task, or

having a normal outcome, for the second, and have values of 0 (negative). A visual representation of the testing structure is shown in Fig 5.2.

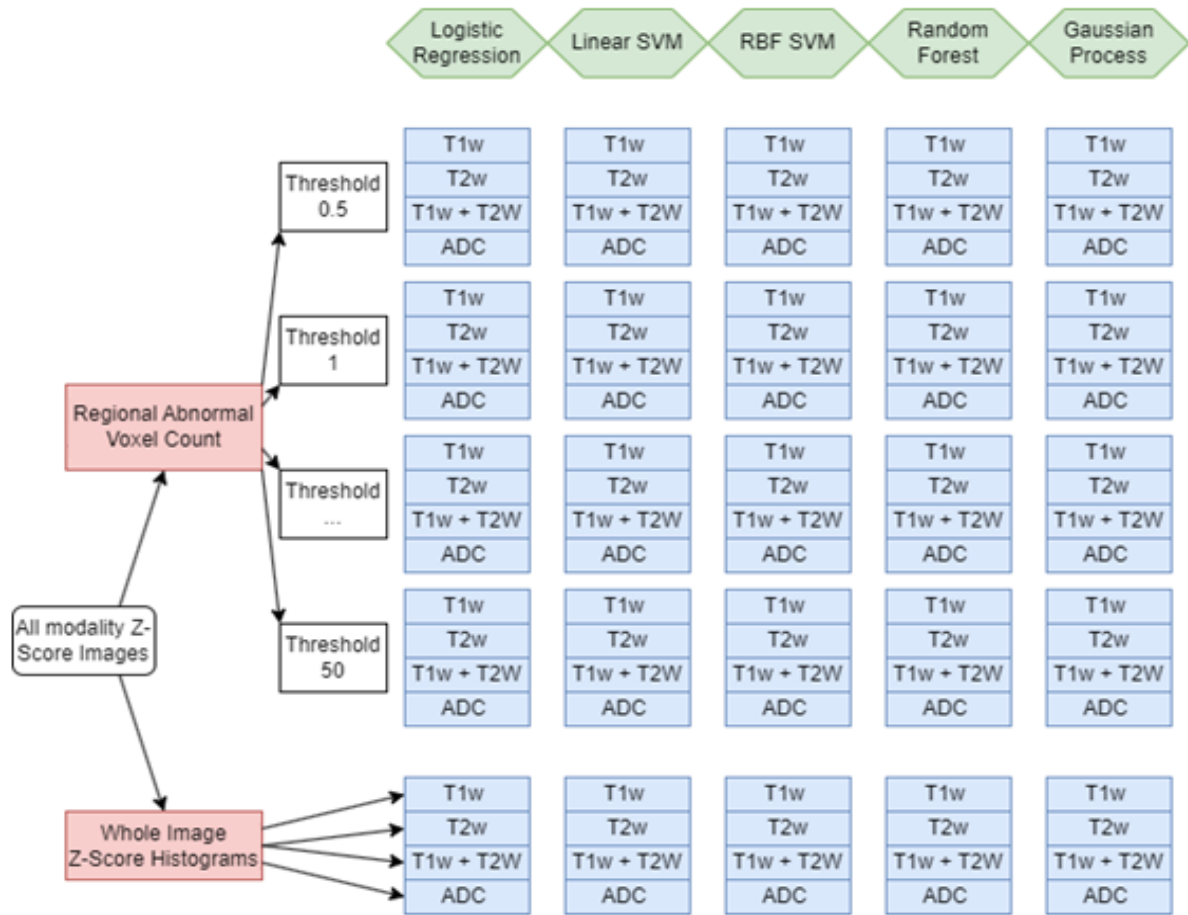


Figure 5.2 Flow chart of the different classifiers used in this chapter. The green hexagons are the different types of classifiers tested. Each classifier constructs 4 models based on inputs from T1w, T2w, T1w+T2w and ADC image Z-Scores. The Z-Score maps have two different ways of extracting data shown in the red rectangles, image-wise histograms of the Z-Scores and the regional abnormal voxel count (RAVC). The RAVC is further subdivided into 100 different models with different thresholds from 0 – 50 in increments of 0.5, white rectangles. The blue rectangles show all the Z-Score histogram models and a cross-section of the RAVC models. All models undergo 5-fold cross-validation. For the 4 best performing Z-Score histogram and RAVC models, a further permutation test is completed.

Additionally, accuracy-based permutation testing was completed on all models. Permutation testing keeps the inputs (histograms or RAVC) the same in all data but randomises the output labels. This gives an indication of what the accuracy would be if the predictions were made for random labels. 1000 permutation tests were used. With this information, the models can be directly compared to random chance, allowing identification of which models had statistically significant performance.

There is an issue with judging performance using such an extensive testing format. This comes in the form of the multiple test comparison problem. As the number of tests or model comparisons increases, the likelihood that one of the models' performances will be significant due to random chance is also increased. To accommodate for this, false discovery rate (FDR) correction was applied to the permutation test p-scores. This is achieved by first ordering the p-scores of each test and applying an index position to each of them. Each initial p-score is multiplied by the total number of tests, 40, divided by the index position of that p-score. This is seen in formula 5.5, where p' is the corrected p-score, p_i is the p score at index i and k is the total number of tests.

$$p' = p_i \frac{k}{i}$$

5.5

5.3 Results

5.3.1 Synthesized Typical Brain Images and Z-scores

Iterative patch-based modelling was able to reproduce the neonatal anatomy in the absence of lesioned tissue within GP produced synthetic images. In areas of tissue containing lesions, accurate representation of normal healthy tissue inferred from the HC dataset was successful, even in the most severe areas of injury, Fig 5.3. As seen in the previous chapter, variance values in abnormal areas for the first iteration of the model were significantly higher than in the

surrounding healthy tissue. Subsequent iterations make predictions conditioned on patches with a more normal appearance, along with GA, PMA and sex, which are more similar to those seen in the HC dataset. This returned the variance values to those seen in healthy tissue while preserving the accurate mean predictions as demonstrated in chapter 4. Higher variance values were observed for areas containing CFS in T2w images, as there is greater variability in intensity.

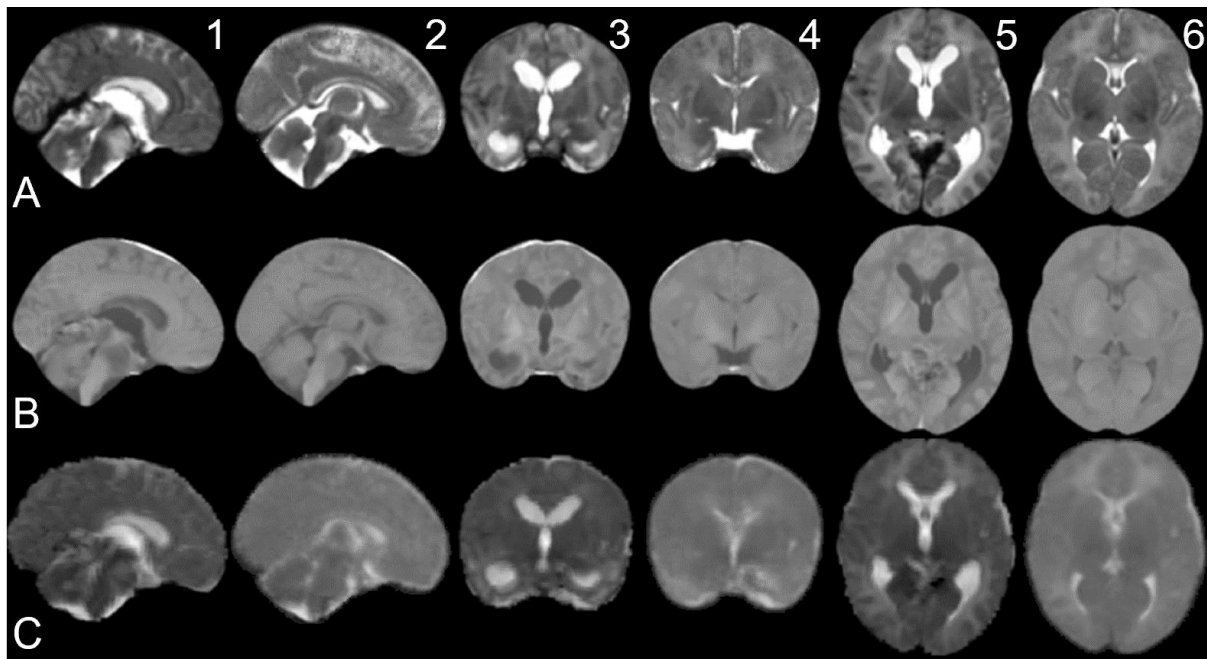


Figure 5.3 Patch-based prediction of HIE neonate with enlarged ventricles. Rows A, B and C show T2w, T1w and ADC images from a HIE neonate with enlarged ventricles. Columns 1, 3 and 5 show the actual MRI images while columns 2, 4 and 6 show the equivalent GP synthesized images.

The Z-score maps created had relatively high values (up to 10) in areas of healthy tissue, caused by the very low variance values in the predictions. However, this is offset by the fact that Z-scores were often very high in areas of abnormality. Values greater than 10 were common and values in the range of 30 to 50 could be observed in certain pathologies (IVH). This is demonstrated in Fig 5.12, which shows raw and Z-score, with thresholds of 5 and 10, overlaid T2w, T1w and ADC images in the sagittal, coronal and axial planes.

5.3.2 Classification of Hypoxic Ischemic Encephalopathy

The HC vs HIE classification results are shown in Figs 5.4 and 5.6. Given the well-balanced nature of the dataset, chance classification should produce base sensitivity and specificity scores of 0.544 and 0.456. In testing, these models demonstrated higher scores, many being above 0.7 and 0.8. The major exception to this was the RBF-SVM models which, except for the ADC based images, failed to converge in the histogram-based models. The best performance, when using Z-score histograms as input, was achieved by the GPC model using ADC images and produced a sensitivity of 0.919 (57/62) and a specificity of 0.942 (47/52). The results of the RAVC threshold testing are shown in table 5.6 which displays each model's threshold and accuracy for T1w and T2w images. The LIN-SVM model using the ADC images gave the highest performance with a sensitivity of 0.935 (58/62) and a specificity of 0.923 (48/52). This model also gave the overall best performance for the separation of HC and HIE neonates. The results of the permutation testing are shown in Figs 5.5 and 5.7 all of which demonstrated statistical significance before and after correction except for the RBF-SVM models.

Table 5.6 Regional abnormal voxel count optimal threshold testing. The first value in each cell is the threshold with the highest accuracy and the second is the accuracy score.

Model	LR	LIN-SVM	RBF-SVM	RF	GPC
T1w Image	2/0.812	3//0.764	5.5/0.778	0.7/0.835	0.3/0.808
T2w Image	3.5/0.822	3.5/0.780	4.5/0.786	0.5/0.821	0.4/0.813

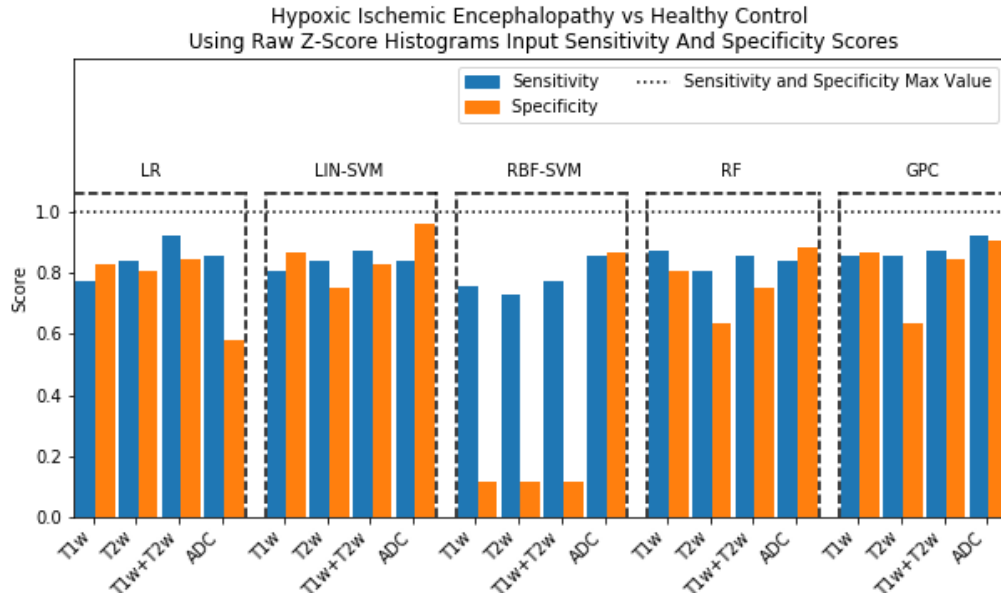


Figure 5.4 HIE vs HC using histograms. Bar plotted sensitivity (blue) and specificity (orange) scores (y-axis) for every model type (dashed lines) and input modality (x-axis). The dotted line at 1 indicates the maximum possible score for each measure.

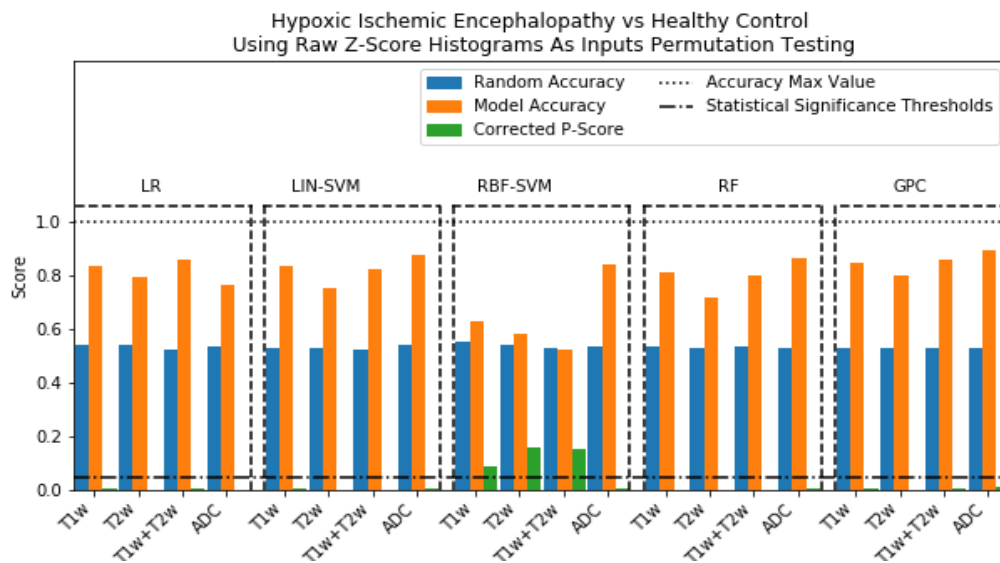


Figure 5.5 HIE vs HC histogram permutation testing. Bar plotted random (blue) and model (orange) accuracy and p-value scores (y-axis) for every model type (dashed lines) and input modality (x-axis). The dotted line at 1 indicates the maximum possible score for each measure and the dotted and dashed line at 0.05 is the statistical significance threshold, if the green bar is larger than this then the results are not significant.

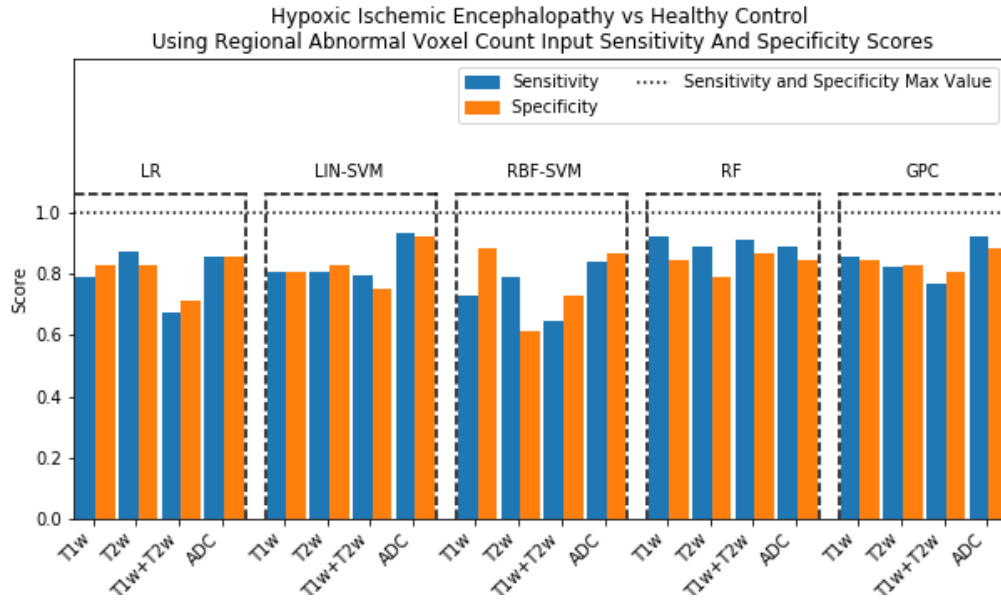


Figure 5.6 HIE vs HC using regional abnormal voxel count. Bar plotted sensitivity (blue) and specificity (orange) scores (y-axis) for every model type (dashed lines) and input modality (x-axis). The dotted line at 1 indicates the maximum possible score for each measure.

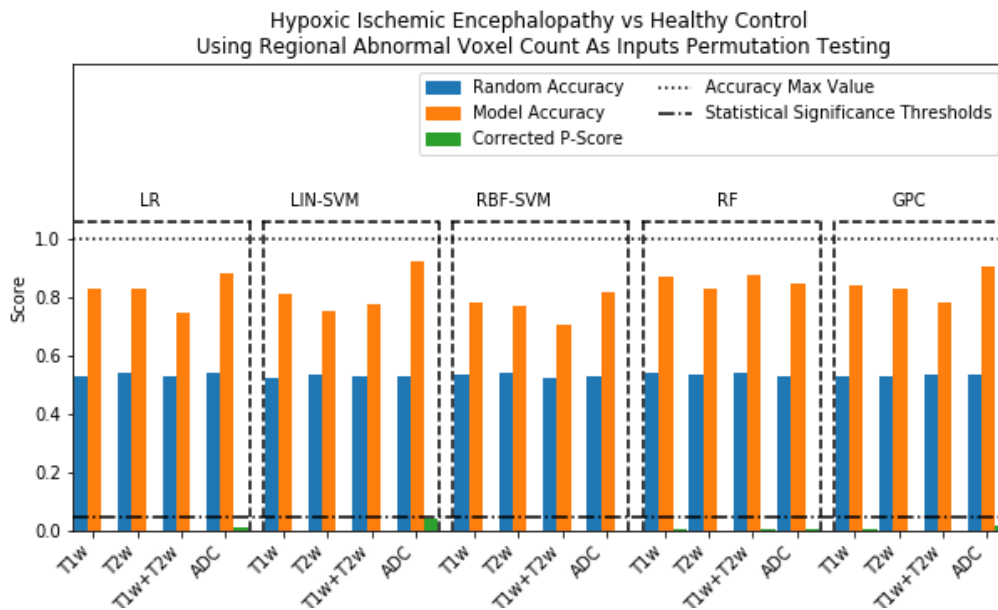


Figure 5.7 HIE vs HC regional abnormal voxel count permutation testing. Bar plotted random (blue) and model (orange) accuracy and p-value scores (y-axis) for every model type (dashed lines) and input modality (x-axis). The dotted line at 1 indicates the maximum possible score for each measure and the dotted and dashed line at 0.05 is the statistical significance threshold, if the green bar is larger than this then the results are not significant.

5.3.3 Classification of Positive and Negative Outcomes

The results of the various classifiers to separate HIE neonates with negative outcome from those with positive (typical) outcomes can be seen in Figure 5.8 and 5.10. For the models using histogram as the input, the best performing was the RF classifier using ADC inputs which achieved a sensitivity of 0.706 (12/17) and a specificity of 0.733 (11/15). For the models using the RAVC, the RF model using only T2w image RAVC gave the best performance with a SENS sensitivity 824 (14/17) and a SPEC specificity 733 (11/15). This model also gave the overall best performance for the separation of HIE neonates with positive and negative outcomes. The results of the permutation testing can be seen in Figures 5.9 and 5.11 and show that after FDR correction the model's predictive ability was not statistically significant.

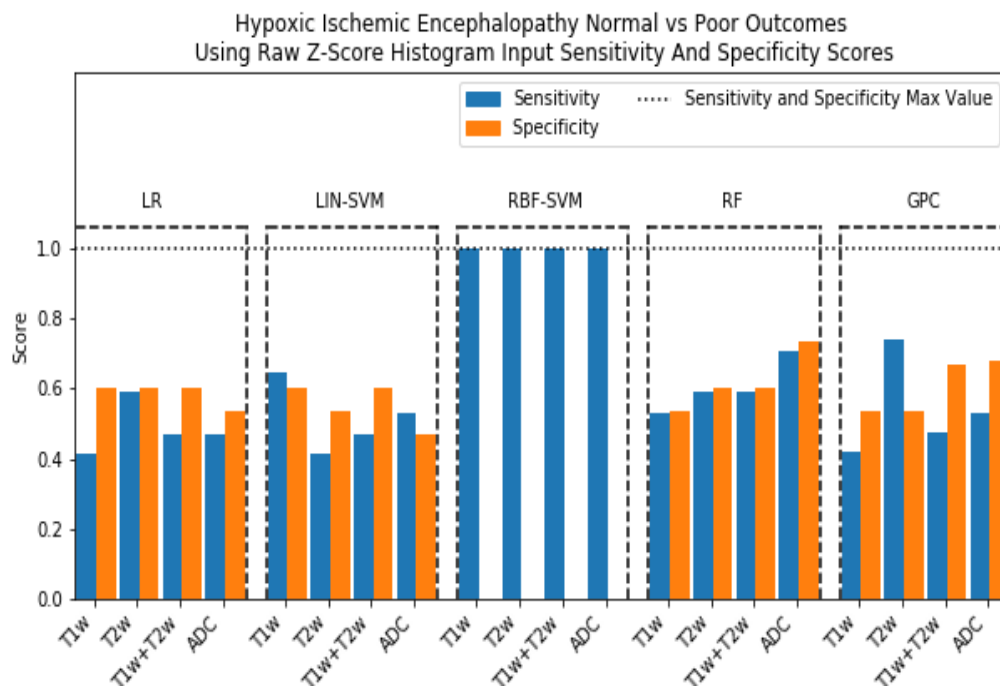


Figure 5.8 HIE normal vs poor outcomes using histograms. Bar plotted sensitivity (blue) and specificity (orange) scores (y-axis) for every model type (dashed lines) and input modality (x-axis). The dotted line at 1 indicates the maximum possible score for each measure.

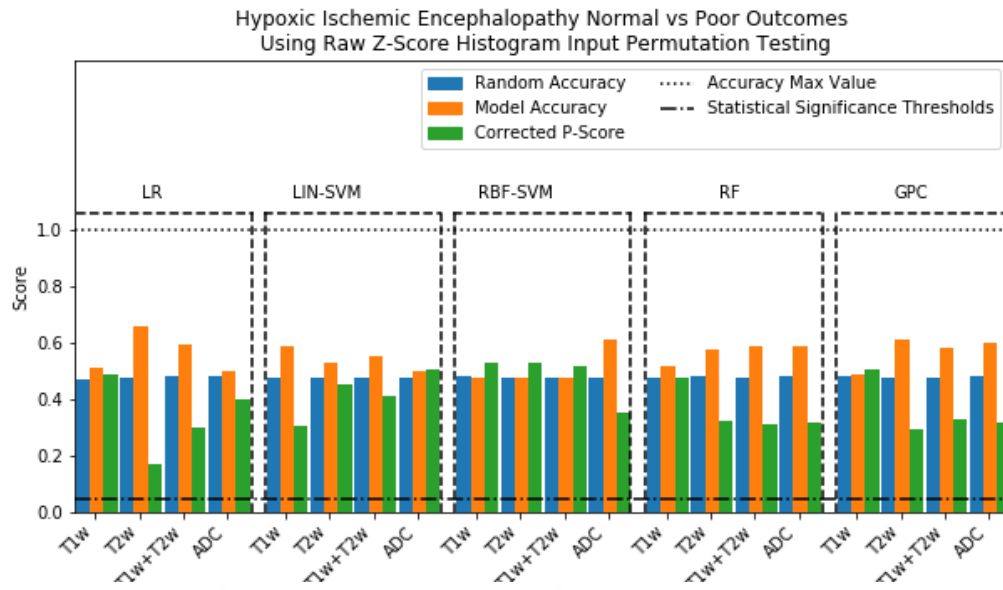


Figure 5.9 HIE normal vs poor outcomes using histograms permutation testing. Bar plotted random (blue) and model (orange) accuracy and p-value scores (y-axis) for every model type (dashed lines) and input modality (x-axis). The dotted line at 1 indicates the maximum possible score for each measure and the dotted and dashed line at 0.05 is the statistical significance threshold, if the green bar is larger than this then the results are not significant.

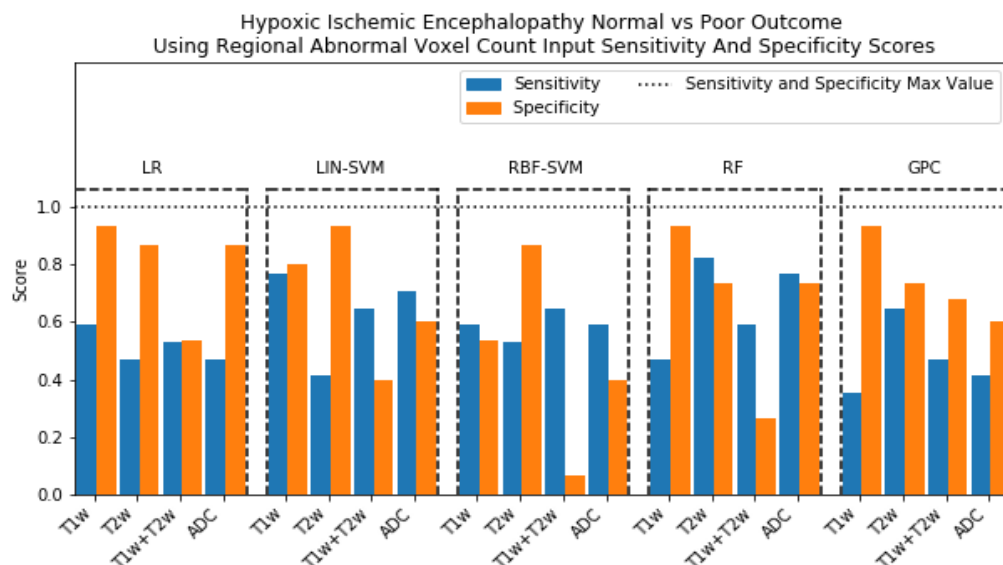


Figure 5.10 HIE normal vs poor outcomes using regional abnormal voxel count. Bar plotted sensitivity (blue) and specificity (orange) scores (y-axis) for every model type (dashed lines) and input modality (x-axis). The dotted line at 1 indicates the maximum possible score for each measure.

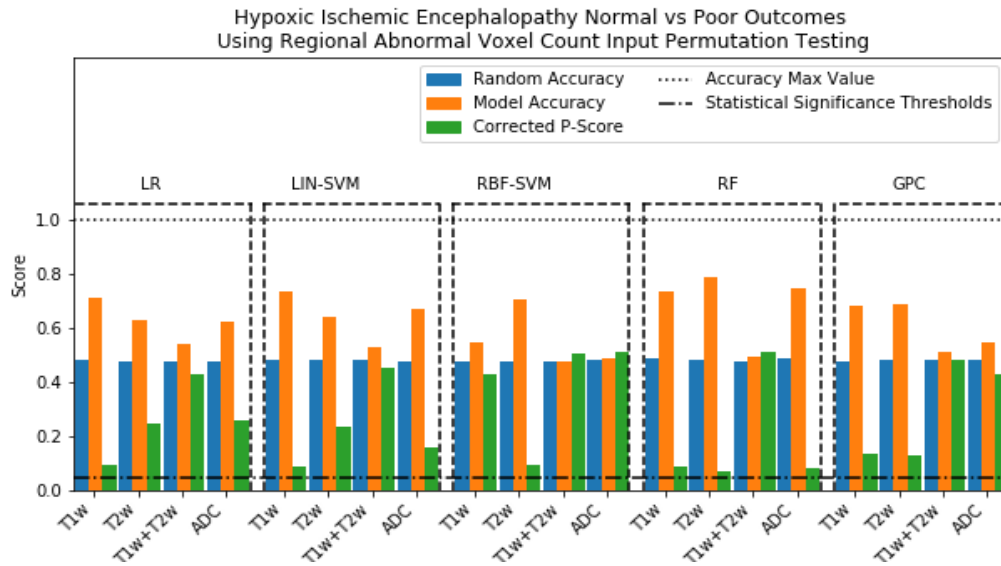


Figure 5.11 HIE normal vs poor outcomes using histograms permutation testing. Bar plotted random (blue) and model (orange) accuracy and p-value scores (y-axis) for every model type (dashed lines) and input modality (x-axis). The dotted line at 1 indicates the maximum possible score for each measure and the dotted and dashed line at 0.05 is the statistical significance threshold, if the green bar is larger than this then the results are not significant.

5.3.4 Identification of Important Features

To determine how the models are making their decisions, an examination of model parameter weightings, optimised during model training, was undertaken. This is possible for the LR model, as it has a directly accessible list of parameter weights. The assessment of the RAVC features demonstrated several areas of the brain that had high weightings across the models. For the HC vs HIE classifiers the following were identified: Left and right hippocampus, right medial part of the anterior temporal lobe GM, left and right gyri anterior part of the parahippocampalis et ambiens GM, left and right cerebellum, left and right thalamus, brainstem and CSF. For the normal vs poor outcomes, the following had consistently high weightings: left and right hippocampus, left and right anterior lobe GM, left and right medial and inferior temporal gyri + posterior gyrus fusiformis GM, left and right thalamus and left and right lentiform nucleus.

5.4 Discussion

This work focused on investigating the use of normative models to classify hypoxic ischemic encephalopathy (HIE) based on clinically acquired multimodal MRI. I trialled both classification and, a more clinically relevant and difficult problem, outcome prediction. This was achieved via a multilayer system. A Z-score map was created using the subject's actual and GP synthesized images, covered in chapters 3 and 4, and corresponding standard deviation map. These Z-score maps were used as inputs into a range of classifiers (LR, LIN-SVM, RBF-SVM, RF, GPC) to either separate HC from HIE neonates or HIE neonates with normal or poor outcomes. In total, 40 models were estimated and tested, five each for combination of T1w, T2w, T1w+T2w and ADC images, for the histogram and RAVC inputs. For the HC vs HIE classification, linear SVM using ADC input had the highest performance (max >90% accuracy). The more challenging outcome prediction had lower performance (max >70% accuracy), and a simple random forest model using T2 inputs performed best.

As expected from the previous results chapter, qualitative assessment of the synthesised normal images demonstrated the model's ability to match heterogeneous brain structure and normal tissue intensity. This approach was successful at detecting injuries more commonly associated with HIE, such as damage in the basal ganglia shown in Figs 5.13-14 (Quattrocchi et al., 2016; Rutherford et al., 2010; Volpe, 2012).

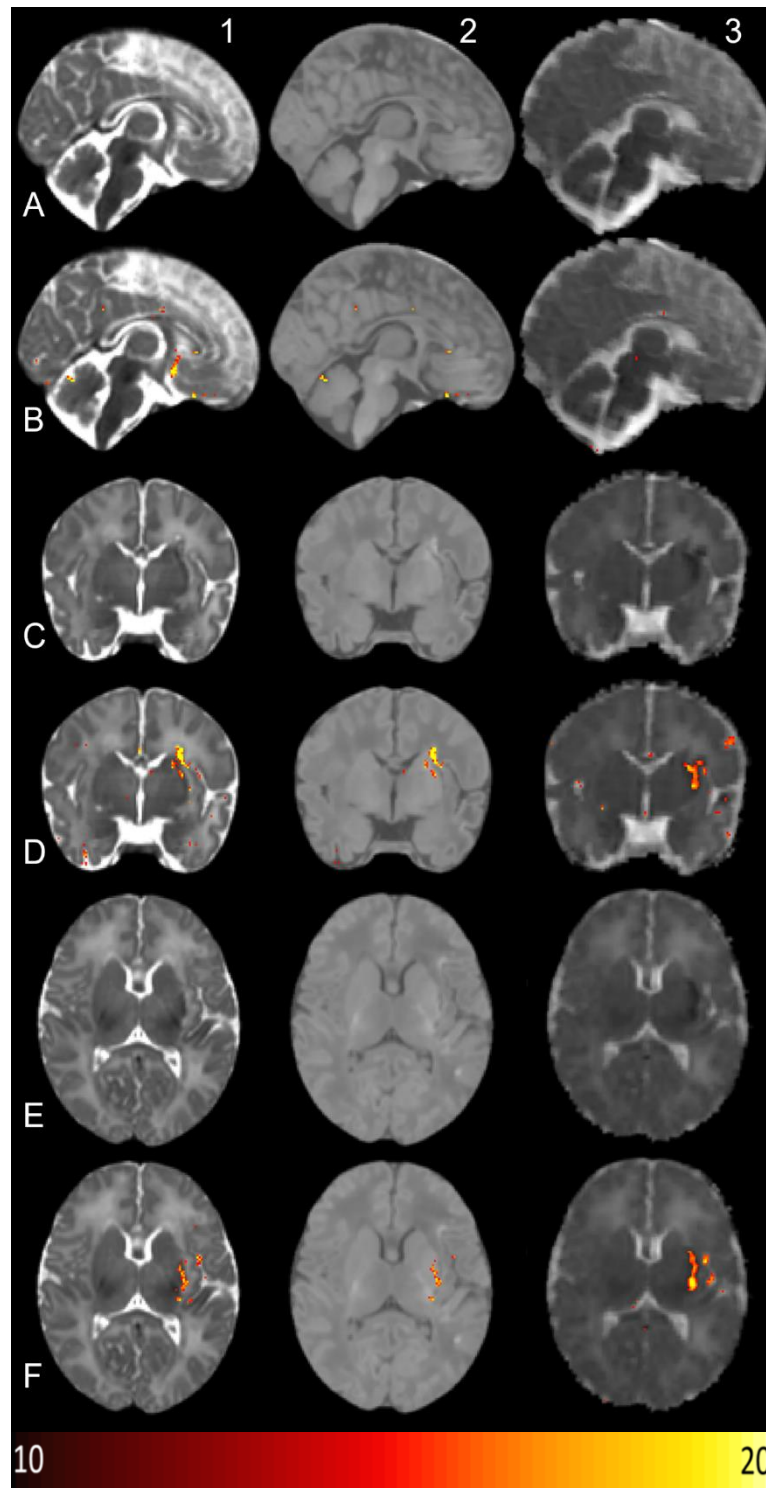


Figure 5.12 Typical HIE lesion detection. Rows A, C and E show a typical HIE neonate's actual T1w (column 1), T2w (column 2) and ADC (column 3) images. Rows B, D and F show the same images with their Z-score overlays with minimum threshold set to 10 and maximum threshold set to 20.

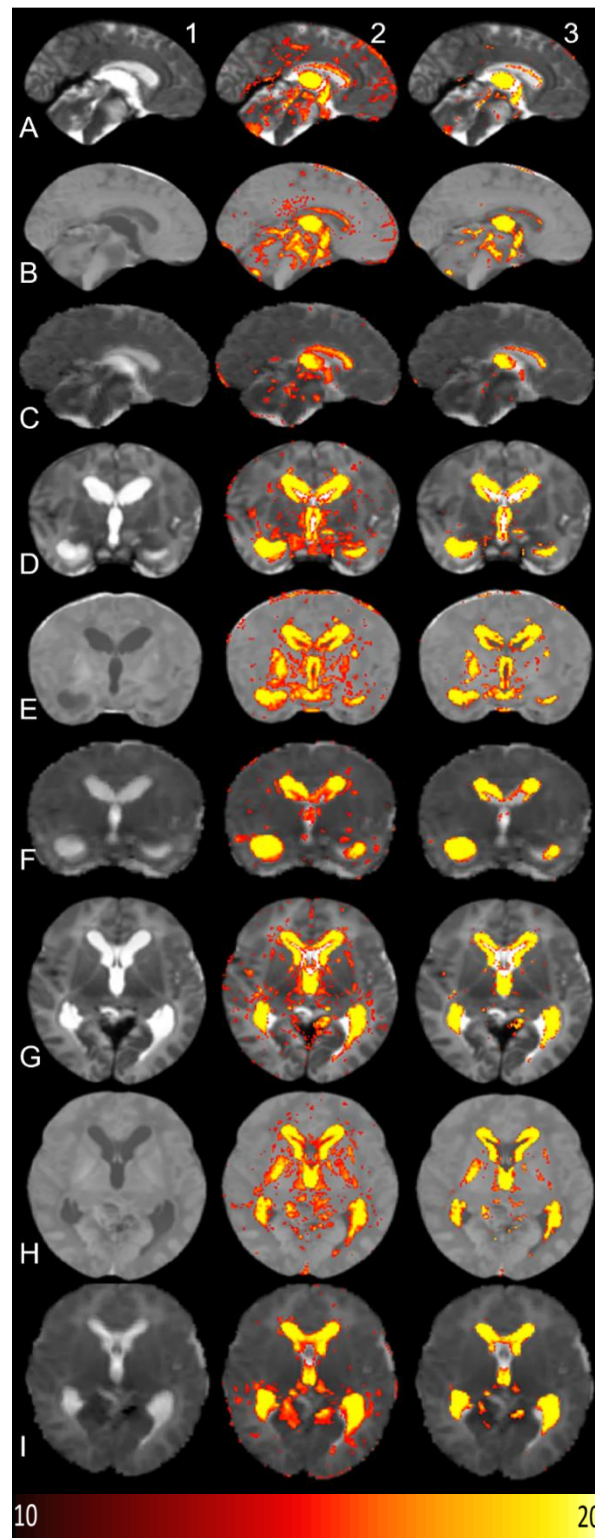


Figure 5.13 Pathology detection at different thresholds. This image shows the same neonates seen in Fig 5.3. Rows A, B and C show the T2w, T1w and ADC images in the sagittal plane, D, E and F, the coronal plane and G, H and I the axial plane. Column 1 shows the raw image, column 2 the same image with the Z-score overlaid with minimum threshold set to 5 and maximum threshold set to 20 and column 3 the overlay with minimum threshold set to 10 and maximum threshold set to 20.

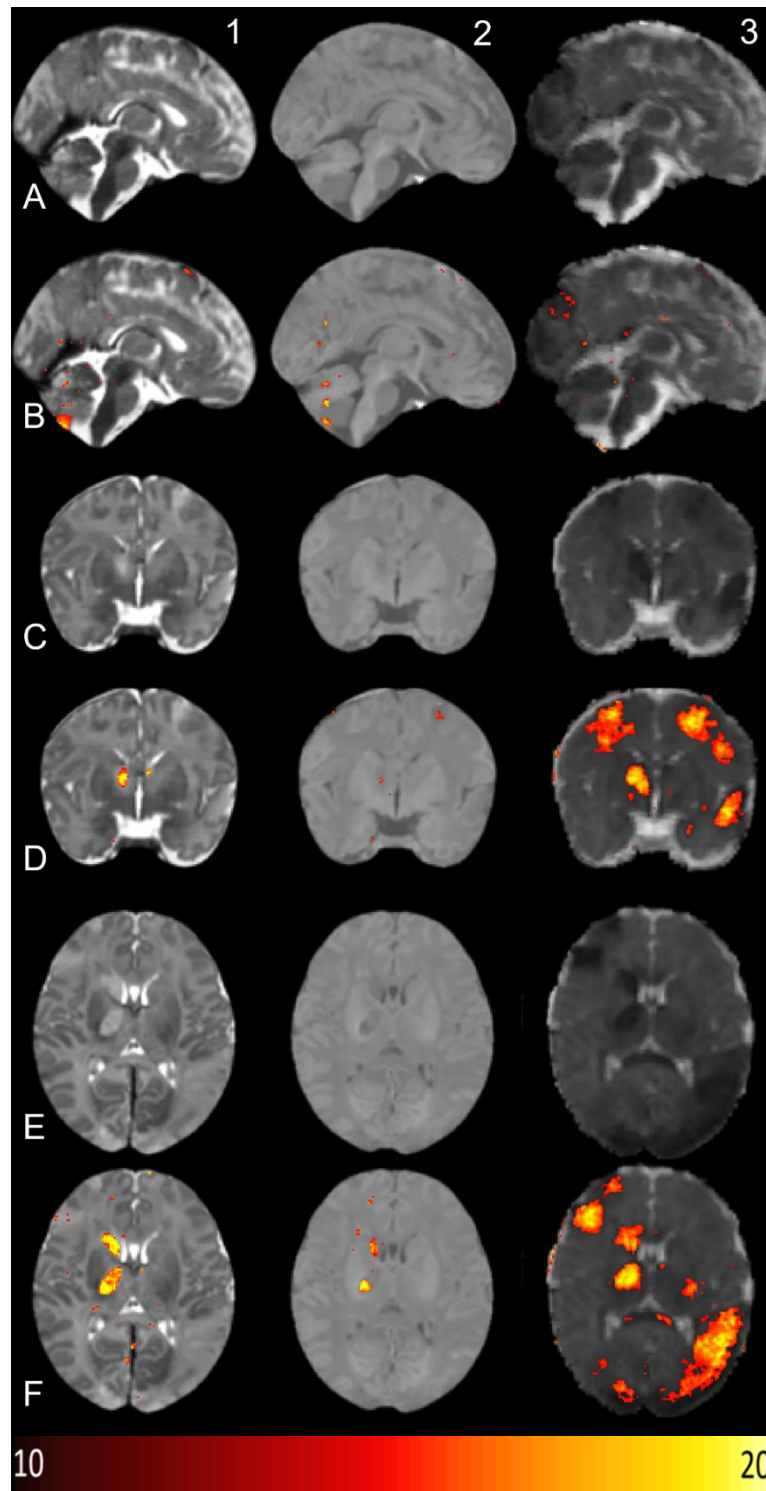


Figure 5.14 Typical HIE lesion detection with emphasis on ADC. Rows A, C and E show a typical HIE neonate's actual T1w (column 1), T2w (column 2) and ADC (column 3) images. Rows B, D and E show the same images with their Z-score overlays with thresholds of 10 to 20. Special attention should be paid to the ADC images and the greater lesion content not visible on other modalities.

The choice of classifier impacted sensitivity, specificity and accuracy. LR and LIN-SVM models showed good performance when applied to the HC vs HIE task for both histogram and RAVC inputs. RF and GPC showed similar results, except for the histogram-based RF and GPC models using T2w inputs which demonstrated marginally lower performance. The poorest models were the RBF-SVM which, with the exception of the ADC model, consistently failed to converge in histogram models, resulting in them predicting all neonates as being HIE. Interestingly the choice of model was also task dependent. LIN-SVM and GPC classifiers had the best scores in the HIE vs HC task, for both histogram and RAVC inputs. RF classifiers showed better scores for HIE normal vs poor outcome tasks.

For HC vs HIE classification task, a LIN-SVM classifier using RAVC inputs from ADC images gave the best performance. Using RAVC based inputs notably improved scores for the RF and RBF-SVM models and all showed marginal improvement when using ADC inputs. All other modalities showed roughly the same level of separation capability. Permutation testing demonstrated that most models were more than capable of outperforming random permutation. The HIE outcome models were unfortunately not able to achieve the same levels of performance. The best performing model was a RF using RAVC inputs from T2w images. Most models showed sensitivity and specificity scores around or just above random chance using histogram inputs, the exception being RF classifiers which were higher. For many of these models, there was a large rise in specificity scores. None of the corrected p-values for the HIE outcome models were statistically significant, although was to be expected given the introductory nature of this work and the small sample size.

This lower performance seen in the normal vs poor outcome models can be attributed to several factors. First, is the fact that it is a more difficult task as the correlation between injury and negative outcome is still a highly debated topic, whereas HIE diagnosis is standardised and overseen by large clinical teams. This is noted by Robertson & Finer, 1985 and Weiss et al.,

2019 and can be seen within this dataset as there exists a neonatal MRI that contains a severe injury with a normal outcome and, conversely, an MRI with a less severe injury that developed negative long-term outcomes seen in Fig 5.15. This is further confounded by the fact that clinical outcome notes were reported from the originating hospitals and inconsistent.

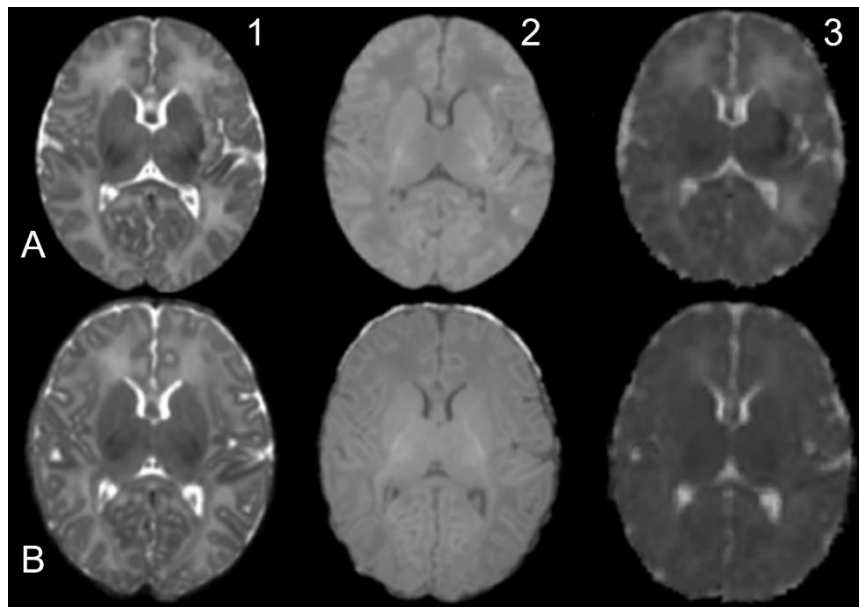


Figure 5.15 Neonates with difficult to predict outcomes. Row A shows a HIE neonate with injury to the basal ganglia and subcortical white matter injury who went on to have normal outcome. Row 2 shows a HIE neonate with a normal appearing brain who had global developmental delay and cerebral palsy. Columns 1, 2 and 3 show T2w, T1w and ADC modality images.

The second factor affecting this task is the relatively small amount of data available. Only 32 HIE neonates with all T1w, T2w, ADC images and patient outcome information were available (15 normal, 17 poor). As such, the classifiers had very limited amount of information with which to construct a model during the cross-validation. It is possible that with a larger, more comprehensive dataset the performance could be improved so this result should be considered preliminary.

There is also the issue of the type and extent of the negative outcomes. Once again due to the size limitation of the dataset, the poor outcome group consisted of all types of negative outcome from limited developmental delay to morbidity. As noted in Rutherford et al., 2010, certain outcomes are often associated with different types of injury, for example damage to the BGT often results in motor delay and cerebral palsy. This limited size and different relationships between injury and outcomes makes inferring the mapping between images and outcome more difficult. This is accentuated by the fact that any outcome may not have a direct relationship to the HIE injury but also due to additional external (for instance body infection or spinal injuries) or genetic factors.

The ability to discriminate between HC and severe HIE cases is expected due to the large quantity of high Z-score voxels observed. What was more interesting was the ability of the various classifiers to correctly identify the more subtle HIE injuries. In many cases, these would have been difficult to detect due to small differences in abnormal tissue intensities and comparatively low Z-score. Despite this the best performing models were able to achieve high levels of sensitivity and specificity. This was supported by the results of the permutation testing which confirmed the equally high, and statistically significant results compared to random permutation after correction.

An interesting facet of this work is the comparative performance of the different imaging modality models. Previous work undertaken in chapter 4 of this thesis (Macleod et al., 2021), found that using T2w images and combining T2w and T1w images both provided greater differentiability between HC and abnormal images than T1w images. In this work T1w images had much better performance, having sensitivity and specificity scores comparative to other modalities. The inclusion of the ADC images is also supported due to its excellent model performance. A reason for this is due to its ability to highlight injury which may not be present in the T1w and T2w images, as seen in Fig 5.14. This is supported by the literature, which

highlight the importance in using multiple modalities when estimating the extent and severity of HIE (Bano et al., 2017; Krishnan & Shroff, 2016; Rutherford et al., 2010; Volpe, 2012).

One advantage of using an LR model is its high interpretability and allowing for examination of its parameter weightings. These showed that bins for the more extreme Z-scores had higher weightings and represents the fact that higher numbers of abnormal voxels, regardless of the direction of intensity difference, are more likely to occur in an injured brain. This makes sense when trying to differentiate between HC, where low numbers of high Z-scores are expected, and HIE neonates, where high numbers are more likely. This LR weighting pattern was repeated for poor vs normal outcome models, although in this instance, its performance was less than optimal. This is likely due to needing more local information about the injury. A large number of differences in the cortical WM may not necessarily result in negative outcomes but a comparatively small number of differences in the BGT may do. The more global score of the histograms lacks the ability to differentiate between these injury types as it does not incorporate positional information.

Several of the areas identified by analysis of the RAVC models match those suggested in HIE injury pattern literature. For the HC vs HIE models, the high weighting of both the cortical and deep GM structures matches the injury patterns suggested in Bano et al., 2017, Rutherford et al., 2010 and Volpe, 2012. The identification of the brainstem and cerebellum also, to a more limited degree, suggested as being important by Lemmon et al., 2017, Quattrocchi et al., 2016 and (Sargent et al., 2004). The identification of CSF is also unsurprising as increased levels are associated with abnormal development and atrophy after damage.

The regions identified in the normal vs poor outcome models also match areas identified as suffering greater amounts of damage in severe HIE and therefore associated with poorer outcomes. Again, the thalamus is identified along with the lentiform nucleus, both members of

the BGT where lesions are known to contribute to severe cases of HIE. Surprisingly, this did not extend to other parts of the BGT such as the caudate and subthalamic nuclei, although this may be due to the limited number of cases. This is also true of areas of the cortical GM and hippocampus (Bobba et al., 2023).

Post modelling analysis of the images provides an interesting insight into the workings of the different models. The histogram-based models consistently identified all of the HIE neonates with obvious pathological lesions while correctly classifying HC neonates, but the RAVC models would sometime miss one particular pathological neonate. This neonate's MRI contained only a subcortical WM haemorrhage and had normal tissue intensities elsewhere in the brain. This is interesting as this type of injury is not by itself strongly associated with HIE (Millar et al., 2017). Further, when compared to the intensity difference seen in this type of injury to those associated with HIE, such as PWML or diffuse WM injury, the intensity shift is in the opposite direction. This suggests that the RAVC models place more importance on features representing voxel intensity differences seen in HIE rather than the more generic global Z-score counts histogram-based models use. This could be an interesting direction to explore in the future; comparing neonates with HIE lesion patterns and neonates with brain injury not representative of HIE and attempting to discriminate between them.

A further issue affecting this work was the limited HC data set size. This was smaller than the normative cohort used in the previous chapters but was applied to a smaller age range, 37-46 compared to 26-45 weeks PMA, and could rely on assistance from the patch-based inputs. The voxel-wise GP regression used in this work needs a reference HC dataset with which to construct its normative models and as a result they are all used for this process. This is done in order to maximise the coverage and inference between input and output variable which is important when trying to accommodate the individual heterogeneity seen in the brain. So, for the HC vs HIE classification task, the only data available were those HC used for the normative

modelling. This means that the generated normal images used in HIE Z-score maps were based on those HC images used in the classification task and represents a potential over use of the dataset (Varoquaux, 2018). A better solution would be to utilise two separate HC datasets. The first for GP model normative cohort and the second as a held-out HC dataset for the classification.

A linked issue is the small cohort size of the HIE neonates with outcome information, which has a limiting effect on the complexity of the classification models. This could be alleviated by gaining access to more data but as infants are referred from other hospitals, chasing outcome data can be extremely challenging. The HIE training data contained many additional neonates who may have outcome information, but which was not accessible for this work and would require follow-up. Alternatively, external HIE data sets with outcome information may be available, such as in Murphy et al., 2017, one of the papers reviewed in Ouwehand et al., 2020 or, when data acquisition is completed, access to the large multi-site cohorts in Weiss et al., 2019. If attempted, image harmonization would be a further challenge, as there are likely to be differences in acquisition parameters, though the patch framework taken for GP models here may act as a potential normaliser across acquisitions.

One direction that would be interesting to explore would be in trying to predict not just if a neonate with HIE is going to have a negative outcome but precisely what type of outcome it would be (speech, motor, cognitive). This has been a focus in adult and neonatal stroke research, though needs very large sample size and detailed outcome follow-up (Hope et al., 2013; MacKay et al., 2020). Additionally, in combination with a probabilistic or ordinal modelling method, it could be extended to prediction of not just the type of negative outcome but also its degree, as seen in Doyle et al., 2013. These improvements have the potential to not only be clinically useful but also allow the models to make more informative mappings

between injury locations and their likely effects on outcome, providing insight on early brain systems underlying cognitive development.

5.5 Conclusion

The results of these experiments demonstrate the efficacy of normative modelling methods used in chapter 4 of this thesis (Macleod et al., 2021) in specific clinical pathology, HIE. It goes beyond injury detection to classification of disease and attempts prognostic outcome. Although accuracy was not as impressive as the HC vs HIE models, the best performing HIE normal vs poor outcome still achieved reasonable sensitivity and specificity values with a limited dataset. Not only could this most likely be improved with more subjects, but it could also be expanded to predict multiple outcome types. For example, it may be possible to expand to multi-class classification to predict motor, sensory or behavioural delay, all of which have been grouped together in this work.

Chapter 6. Thesis Discussion and Conclusions

The brain develops rapidly over the perinatal period. The gross and subtle changes in anatomy occurring over this timeframe can make identifying clinically abnormal tissue challenging. The goal of my thesis was to build on advances in interpretable machine learning methods and produce an automated means of both (a) identifying neonates with heterogeneous brain structural abnormalities and (b) providing an individualised estimation of these injury locations.

6.1 Thesis Overview

The first section of this thesis covered the implementation of a voxel-wise Gaussian process-based (GP) estimation of typical brain tissue intensity for a given MRI acquisition modality over the perinatal period. This was used to create outlier maps (interpreted as abnormality maps) that demonstrated deviations between GP model predicted values and a neonate's observed MRI values. I then investigated incorporating local tissue anatomy, allowing models to accommodate and predict subject specific heterogeneous brain structure. At this stage, a proof of concept of across subject classification was developed to separate neonates with normal appearing brains from those with abnormality detectable on MRI. The final section of this thesis covered the application of these methods towards a serious brain disease seen in neonates born around typical term age; hypoxic ischemic encephalopathy (HIE). This section also included development of the classification methodology to detect HIE and predict an individual child's outcome. It also explored how information could be extracted from abnormality maps, different methods of classification and attempted to isolate neonates that were likely to develop long-term negative outcome (cognitive or motor).

My initial implementation of an abnormality detection task utilised a voxel-wise GP based normative modelling framework (Rasmussen & Williams, 2005). GPs make predictions based

on the output variable of the normative cohort samples whose inputs most closely match those of the prediction target. More intuitively, data points with similar inputs are likely to have similar outputs. Alongside this mean output, a GP also produces an additional variance characteristic. This incorporates both the model certainty, which increases as the prediction target's inputs become closer to those observed in the normative cohort, and an estimation of the noise present within that normative cohort's outputs. The final useful ability of a GP is its ability to automatically regularise its predictions. If it encounters inputs which are sufficiently disparate from those observed in the training data, it reverts to an optimised prior based on the mean of the normative cohort's outputs.

The GP approach used in this thesis was first demonstrated in Ziegler et al., 2014 for identification of lesions associated with Alzheimer's disease and later by O'Muircheartaigh et al., 2020 for detection of neonatal punctate white matter lesions. Each brain voxel had its own GP model, accessing the same normative cohort (neonates with no apparent injury) and the same set of demographic information including gestational age at birth (GA), post menstrual age at scan (PMA) and sex. When making a prediction, the complete set of GP models output the predicted intensity values for each voxel. These were combined by spatial location to construct a synthesised "normal" appearing brain alongside a corresponding standard deviation map derived from the variance outputs.

The actual abnormality detection was based on construction of a subject specific Z-score abnormality map. This subtracted the neonates actual MRI intensity values from the synthetic image then scaled by the confidence interval of the prediction. Larger values represented greater amounts of deviation between the neonate's MRI and the predicted typical values. A major benefit of scaling by the standard deviation derived from a GP output is that it automatically incorporates the spatially specific intensity variance observed within a given voxel. For example, with a GP modelling a voxel located within the deep grey matter, the

variance was relatively small, as each neonate in the normative training cohort had relatively similar intensity values. For a GP located in a more variable location, such as a tissue boundary or the cortical surface, the variance was larger, as it needed to account for multiple tissue types.

The major contribution of this chapter was the application and estimation of efficacy of this method when used to detect multiple types of common neonatal brain lesion. These pathologies included haemorrhage, cysts, pseudocysts, periventricular leukomalacia, punctate white matter lesions, and general white matter and deep grey matter injury. It also included larger more widespread damage such as ventriculomegaly and severe lesions throughout the basal ganglia and subcortical white matter. A subset of 36 pathological neonates with a selection of pathologies underwent manual segmentation by two independent raters. The accuracy of GP generated Z-score abnormality maps to delineate the manual segmentation images, yielded excellent AUROC scores of 0.961 and 0.947.

However, these values were in part driven by the unbalanced nature of abnormality detection in a neuroimaging context. Unless the neonate has undergone severe trauma, resulting in significant damage to the brain, the majority of values will be representative of normal tissue. This resulted in large numbers of true negatives, which contribute to specificity, overpowering the effect of low numbers of true positives which contribute to the sensitivity score. A large quantity of noisy false positives must be present before having a significant negative effect on the specificity.

Sensitivity scores appear to be relatively low, but this is a consequence of averaging over all neonatal sensitivity scores. Individual sensitivity scores ranged from 0.967 to 0, indicating that in certain instances the model performed exceedingly well but in others couldn't detect a lesion at all. The scores of 0 (seen in 5 infants) could be attributed to specific characteristic of the injuries. For example, punctate lesions could be only 2 to 8 voxels in size and have intensity

values close to those seen in normal tissue. Larger injuries, such as cystic periventricular leukomalacia and germinal matrix haemorrhage, achieved higher scores and showed excellent lesion coverage.

An issue with this initial model was that predictions based purely on demographic information resulted in synthesised images that were grossly appropriate but blurry in areas of high individual heterogeneity (e.g. normal variability of cortical folding patterns). The result was that in these highly variable areas a large quantity of blurring could be observed as the model attempted to accommodate multiple tissue types. This could result in two unwelcome scenarios; the obscuring of pathological tissue which contained intensities within the expected normal range or the identification of normal structure as being pathological. To alleviate these issues, a method of incorporating individual structural heterogeneity into the model would be needed. This could result in the production of synthesised images specific to an individual rather than the more generic images based purely on demographics.

Multiple approaches have been taken to incorporate structural information into abnormality detection models. These include use of graphical models (Markov random fields) seen in N. K. Subbanna et al., 2019, similarity measures (K nearest neighbour, non-local mean) used by Guizard et al., 2015 and deep learning architectures used by Liu et al., 2020. Of particular interest is the method used in Guizard et al., 2015 in which spatially identical patches from a healthy normative cohort are compared to a suspected pathological subject. This method of identification via similarity measures closely resembles that utilised by a GP. Voxel intensities in a patch surrounding the prediction target were incorporated into the GP design matrix alongside the demographic information. This allowed estimation of target voxel intensity to be based on those neonates within the normative cohort whose patches most closely matched that of the neonate being predicted. This allowed a predicted image to match a neonate's

heterogeneous brain structure more closely than one based purely off demographic information.

There is an issue with incorporating patches of local tissue into the model. Patches for a single location within the normative cohort were, besides some small amounts of individual heterogeneity, relatively consistent, which was also the case in areas of healthy appearing tissue of the prediction target. However, in areas containing pathological tissue, this was no longer held true, as abnormal tissue intensities were present in the patch. For the mean prediction, this was not an issue as, when samples are extremely dissimilar, the GPs reverts to the optimised prior mean. The issue arose in the GP variance output. As this included an estimation of model certainty, it increases drastically in patches that were grossly dissimilar from those observed in the normative cohort. This caused a huge increase in the difference needed between an MRI and synthetic voxel before it could be deemed “abnormal”, essentially masking any true abnormalities present.

To correct for this increased variance in abnormal areas, an iterative implementation of the model was created. The first iteration of the model proceeded as usual, with creation of a synthesised normal image and variance map from this patch framework. At this point the model is iterated again, this time using patches from the synthesised image rather than the actual MRI. From the second, and for all subsequent iterations, the GP variance outputs returned to values found in patches composed of normal tissue. While the second and third iterations produced predicted images almost identical to the initial prediction, iteration beyond this point resulted in a gradual degradation of image quality. Essentially, this allowed the first iteration of the model to act as a form of lesion infilling, where voxels containing suspected pathological tissue were replaced with intensities more representative of normal tissue.

To maintain consistency, this new iterative, patch-based approach was tested on the same preterm dataset as used in results chapter one, alongside an additional brain injury data set, twenty of which had manual segmentations. The best performance was found using a 7x7x7 surface patch based on the mean of the T1w and T2w absolute Z-score maps, achieving an AUC score of 0.961. A more in-depth analysis of the sensitivity and specificity scores at specific thresholds supported this, as it consistently maintained higher scores than other patches. This was further reinforced after application to the brain injury dataset, which demonstrated similar results.

One important factor is that while patch-based model lesion coverage was lower in some of the larger pathologies, due to inclusion of lesioned tissue into the synthetic image, it was higher in the smaller more subtle pathologies. Better coverage could also be observed around lesion edges as the subtle intensity differences could be more readily detected by the more sensitive patch-based model. The same selection of very small lesions with minor intensity differences (subtle isolated punctates) were missed in the patch-based model as with the base model. However, lowering the threshold to the point of their detection resulted in fewer false positives in the patch-based model compared to the base model. It can be argued that this is a beneficial trade off. While some of the larger lesions missed the core of their segmentations, they were still detected. In exchange for this, smaller lesions, which may still affect a neonate's health, were detected more consistently.

This chapter also saw a proof-of-concept implementation of across-subject classification between neonates with normal appearing brains and those with lesions via use of a logistic regression classifier. 33 normal appearing neonates, stratified by age, were used as a testing set alongside the 36 pathological images with manual segmentations. All remaining images, normal appearing and with lesions, were used as a training set. Combining the predictive power of separate T1w and T2w classifiers increased accuracy, correctly identifying 32/36 images

containing pathological lesions and 28/33 neonates with no obvious lesions. This performance suggests that the GP generated abnormality maps have additional uses as inputs to secondary machine learning applications or for imaging triage. The fact that better performance was seen in models that utilised multi-modality inputs is supported by statements in clinical literature, many of which indicate that certain acquisitions are better at highlighting different injuries (Rutherford, Biarge, et al., 2010).

The final focus was on application of these techniques to a specific pathological case, hypoxic ischemic encephalography (HIE), as well as further enhancements to the classification method. The first of these enhancements involved use of alternative classifiers, including random forest, support vector machine (linear kernel, radial basis function kernel) and a Gaussian process classifier. It also investigated the use of more localised information, counting the number of voxels with Z-scores over a given threshold within different regions, instead of the more global Z-score histograms used previously. To ensure robustness, extensive permutation testing was applied to all models. This first constructs the model normally and estimates its accuracy via 5-fold cross validation. It then fixes the output labels but randomises the input variables and attempts to reconstruct the classifier on this randomised data, a process repeated 1000 times. The output is a p-value that indicates the number of times a model built on the random data achieved an accuracy greater than that of the original model. Due to the large number of models, it is possible that one could produce high levels of accuracy and p-values by random chance. To compensate for this, false discovery rate correction was applied to the p-values to ensure their statistical significance.

Multiple combinations of model type, input type, an image modality were compared to assess performance when trying to separate HIE and healthy control neonates. This was achieved by an SVM model using regional scores from ADC Z-score images. It correctly identified 58/62 HIE neonates and 48/52 healthy neonates, demonstrating accuracy over 90%. The majority of

these models were also able to demonstrate statistically significant permutation p-values even after correction. This was followed by the more challenging task of predicting long term functional outcome, either normal or poor, given a confirmed HIE diagnosis. This model's performance was not as good as that seen in the healthy control vs HIE classification task. The best performing model, an RF model using regional input from T2w images, was able to correctly identifying 14/17 neonates with poor outcomes and 11/15 neonates with typical outcome and respectable accuracy score >70%. Unfortunately, the p-values generated by the permutation testing did not remain statistically significant after correction.

6.2 Limitations and Improvement

The likely contributor to this poor performance, is the extremely limited size of the dataset available for the task. The original HIE dataset was composed of 114 neonates. However, only 32 neonates had all 3 of the required artefact free images (T1w, T2w, ADC) as well as outcome information. This can be attributed to the fact that this work was essentially a retrospective study, utilising already existing data. A particular issue for the HIE data was that many of these neonates were referred to the imaging hospital from other medical institutions due to its specialised neonatal intensive care unit. While this increased the gross number of subjects, it made accessing follow up information after being discharged more difficult to ascertain. This issue of limited cohort size and data availability could be alleviated by expanding this retrospective study to encompass multiple sites, like that being investigated by Weiss et al., 2019. Alternatively, a (large) prospective study focused HIE, and with an emphasis on long-term follow-up, could be designed.

A larger study could allow a more expansive investigation into outcome prediction. At present, all negative outcomes are grouped together, a less than ideal situation as these can have different aetiologies even within general HIE injury (Bano et al., 2017; Perez et al., 2013). A more expansive dataset, with better coverage of both initial injury and long-term outcomes

could allow a more precise model, potentially able to specify type and/or severity of negative outcome and allowing better interpretation of regional injury and how it contributes to e.g. language outcome. Alternatively, this could be expanded to classification of multiple neonatal pathology types. In keeping with the theme of this thesis, it could allow the classifier to act as a more general form of diagnostic tool, going beyond simply separating healthy controls from either general pathology or HIE. It could also address the issue of accurate performance estimation in small dataset size, more specifically the fact that small testing sets frequently result in large errors when estimating accuracy (Varoquaux, 2018). Although permutation testing can address this bias in terms of accurate estimates of statistically significant, the accuracy value itself can still be imprecise.

The GP models could also be improved and GPs themselves are being rapidly being improved and generalised to new tasks. A basic challenge is thresholding. Successful labelling of lesioned tissue is dependent on the selection of an appropriate Z-score threshold. While something to this effect was achieved in chapter 5 to create the regional scores, a more comprehensive investigation could provide a better selection of optimal thresholds, for instance one specific to a given pathology. Regardless of what this threshold is, it involves an inherent trade-off between maximising lesion coverage while preventing false positives in healthy tissue. This leads neatly into another potential improvement, a method that removes false positives from the initial lesion labels such as that seen in Subbanna et al., 2019, Saladi & Amutha Prabha, 2018 and Roy et al., 2015. Another possibility could be further refinements to the patch-based models. Surface patches were preferred as they are less likely to incorporate pathological tissue into the prediction but are less capable at inferring heterogeneous tissue. Some method of dynamically reducing the contribution of abnormal tissue at inference time would be an ideal balance, as it would allow the use of dense patches, which better estimate typical tissue, whilst minimising the inclusion of pathological tissue. One possibility is an alternative RBF kernel

implementation used by the GP models. This method, called automatic relevance determination (ARD), allows the lengthscale, the similarity measure and therefore contribution, of each input to be different.

6.3 Final Remarks

These abnormality maps have the potential to be applied in a much wider setting, more generally to older children and adults. They could also be used to help identify medical conditions that are either rare or difficult to detect without specialised training as long as they have some form of MRI positive feature. Although only briefly touched on in this work due to its limited applicability, deep learning-based methods of pathology classification are an ever-increasing focus in medical research. These methods rely on isolation of features strongly associated with a given pathology and require incredibly large dataset for effective optimisation. A deep learning-based classifier that incorporates the abnormality maps produced by this work has the potential to outperform both the methods used here and those directly applied to MRI images.

This thesis has focused on investigating the use of the normative modelling framework for the purpose of detecting generic tissue injury and abnormalities in neonates seen in a hospital context. A major benefit of this chosen method, Gaussian Processes, is its good performance when applied to the small cohorts commonly found in the clinical neuroimaging setting. Additionally, and unlike many of the methods seen in the wider pathology detection literature, the method is generic. The generated lesion labels were capable of identifying multiple types of neonatal neurological injuries, including haemorrhages, infarcts, cystic tissue, white and grey matter lesions. These lesion labels have also been used as inputs into multiple classifiers. These were capable of separating neonates with normal appearing brains from those with brain lesions, healthy control neonates from those with a HIE diagnosis and, to a more limited degree, HIE neonates with poor and typical outcome. This demonstrates how these methods could be

used to assist the radiological review process at multiple points, triage, review, diagnosis and prognosis. The machine learning community has focused on larger, common adult diseases (e.g. stroke, multiple sclerosis, Alzheimer's) but this has led to the relative exclusion of rare diseases and paediatrics from taking advantage of the obvious advances. My work aimed to address this imbalance and open the potential of research innovations to the neonatal context.

Chapter 7. Bibliography and Citations

- 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. <https://doi.org/10.1016/J.CLINPH.2015.05.024>
- Alickovic, E., & Subasi, A. (2020). Automatic Detection of Alzheimer Disease Based on Histogram and Random Forest. *IFMBE Proceedings*, 73, 91–96. https://doi.org/10.1007/978-3-030-17971-7_14
- Aljabar, P., Heckemann, R. A., Hammers, A., Hajnal, J. V., & Rueckert, D. (2009). Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *NeuroImage*, 46(3), 726–738. <https://doi.org/10.1016/J.NEUROIMAGE.2009.02.018>
- Anand, K. J. S., & Scalzo, F. M. (2000). Can Adverse Neonatal Experiences Alter Brain Development and Subsequent Behavior? *Neonatology*, 77(2), 69–82. <https://doi.org/10.1159/000014197>
- Andersson, J. L. R., & Sotiropoulos, S. N. (2016). An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *NeuroImage*, 125, 1063–1078. <https://doi.org/10.1016/J.NEUROIMAGE.2015.10.019>
- Avants, B. B., Tustison, N., & Johnson, H. (2014). *Advanced Normalization Tools (ANTs) Release 2.x*. <https://brianavants.wordpress.com/2012/04/13/updated-ants-compile-instructions-april-12-2012/>
- Azami, M. El, Hammers, A., Jung, J., Costes, N., Bouet, R., & Lartizien, C. (2016). Detection of Lesions Underlying Intractable Epilepsy on T1-Weighted MRI as an Outlier Detection

- Problem. *PLOS ONE*, 11(9), e0161498.
<https://doi.org/10.1371/JOURNAL.PONE.0161498>
- Back, S. A. (2014). Cerebral white and gray matter injury in newborns: New insights into pathophysiology and management. *Clinics in Perinatology*, 41(1), 1.
<https://doi.org/10.1016/J.CLP.2013.11.001>
- Back, S. A. (2017). White matter injury in the preterm infant: pathology and mechanisms. *Acta Neuropathologica* 2017 134:3, 134(3), 331–349. <https://doi.org/10.1007/S00401-017-1718-6>
- Bahadure, N. B., Ray, A. K., & Thethi, H. P. (2017). Image Analysis for MRI Based Brain Tumor Detection and Feature Extraction Using Biologically Inspired BWT and SVM. *International Journal of Biomedical Imaging*, 2017.
<https://doi.org/10.1155/2017/9749108>
- Ballabh, P. (2010). Intraventricular Hemorrhage in Premature Infants: Mechanism of Disease. *Pediatric Research* 2010 67:1, 67(1), 1–8.
<https://doi.org/10.1203/pdr.0b013e3181c1b176>
- Bano, S., Chaudhary, V., & Garga, U. C. (2017). Neonatal Hypoxic-ischemic Encephalopathy: A Radiological Review. *Journal of Pediatric Neurosciences*, 12(1), 1.
<https://doi.org/10.4103/1817-1745.205646>
- Barkovich, A. J. (2000). Concepts of Myelin and Myelination in Neuroradiology. *AJNR Am J Neuroradiol*, 21, 1099–1109.
- Barkovich, A. J. (1998). MR of the normal neonatal brain: assessment of deep structures. *American Journal of Neuroradiology*, 19(8).

- Bassan, H., Feldman, H. A., Limperopoulos, C., Benson, C. B., Ringer, S. A., Veracruz, E., Soul, J. S., Volpe, J. J., & Plessis, A. J. du. (2006). Periventricular Hemorrhagic Infarction: Risk Factors and Neonatal Outcome. *Pediatric Neurology*, 35(2), 85–92. <https://doi.org/10.1016/J.PEDIATRNEUROL.2006.03.005>
- Batalle, D., Edwards, A. D., & O’Muircheartaigh, J. (2018). Annual Research Review: Not just a small adult brain: understanding later neurodevelopment through imaging the neonatal brain. *Journal of Child Psychology and Psychiatry*, 59(4), 350–371. <https://doi.org/10.1111/JCPP.12838>
- Bats, A. S., Molho, M., Senat, M. V, Paupe, A., Bernard, J. P., & Ville, Y. (2002). *Subependymal pseudocysts in the fetal brain: prenatal diagnosis of two cases and review of the literature.*
- Baur, C., Denner, S., Wiestler, B., Navab, N., & Albarqouni, S. (2021). Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study. *Medical Image Analysis*, 69, 101952. <https://doi.org/10.1016/J.MEDIA.2020.101952>
- Bethlehem, R. A. I., Seidlitz, J., White, S. R., Vogel, J. W., Anderson, K. M., Adamson, C., Adler, S., Alexopoulos, G. S., Anagnostou, E., Areces-Gonzalez, A., Astle, D. E., Auyeung, B., Ayub, M., Bae, J., Ball, G., Baron-Cohen, S., Beare, R., Bedford, S. A., Benegal, V., ... Alexander-Bloch, A. F. (2022). Brain charts for the human lifespan. *Nature* 2022 604:7906, 604(7906), 525–533. <https://doi.org/10.1038/s41586-022-04554-y>
- Boardman, J. P., & Counsell, S. J. (2020). Invited Review: Factors associated with atypical brain development in preterm infants: insights from magnetic resonance imaging. *Neuropathology and Applied Neurobiology*, 46(5), 413–421. <https://doi.org/10.1111/NAN.12589>

- Boardman, J. P., Counsell, S. J., Rueckert, D., Kapellou, O., Bhatia, K. K., Aljabar, P., Hajnal, J., Allsop, J. M., Rutherford, M. A., & Edwards, A. D. (2006). Abnormal deep grey matter development following preterm birth detected using deformation-based morphometry. *NeuroImage*, 32(1), 70–78. <https://doi.org/10.1016/J.NEUROIMAGE.2006.03.029>
- Boardman, J. P., Walley, A., Ball, G., Takousis, P., Krishnan, M. L., Hughes-Carre, L., Aljabar, P., Serag, A., King, C., Merchant, N., Srinivasan, L., Froguel, P., Hajnal, J., Rueckert, D., Counsell, S., & Edwards, A. D. (2014). Common Genetic Variants and Risk of Brain Injury After Preterm Birth. *Pediatrics*, 133(6), e1655–e1663. <https://doi.org/10.1542/PEDS.2013-3011>
- Bobba, P. S., Malhotra, A., Sheth, K. N., Taylor, S. N., Ment, L. R., & Payabvash, S. (2023). Brain injury patterns in hypoxic ischemic encephalopathy of term neonates. *Journal of Neuroimaging*, 33(1), 79–84. <https://doi.org/10.1111/JON.13052>
- Bogner, M. S. (2018). Human error in medicine. In *Human Error in Medicine*. <https://doi.org/10.1201/9780203751725>
- Bowles, C., Qin, C., Guerrero, R., Gunn, R., Hammers, A., Dickie, D. A., Valdés Hernández, M., Wardlaw, J., & Rueckert, D. (2017). Brain lesion segmentation through image synthesis and outlier detection. *NeuroImage: Clinical*, 16, 643–658. <https://doi.org/10.1016/J.NICL.2017.09.003>
- Bozek, J., Makropoulos, A., Schuh, A., Fitzgibbon, S., Wright, R., Glasser, M. F., Coalson, T. S., O’Muircheartaigh, J., Hutter, J., Price, A. N., Cordero-Grande, L., Teixeira, R. P. A. G., Hughes, E., Tusor, N., Baruteau, K. P., Rutherford, M. A., Edwards, A. D., Hajnal, J. V., Smith, S. M., ... Robinson, E. C. (2018). Construction of a neonatal cortical surface

- atlas using Multimodal Surface Matching in the Developing Human Connectome Project. *NeuroImage*, 179, 11–29. <https://doi.org/10.1016/J.NEUROIMAGE.2018.06.018>
- Branson, H. M. (2013). Normal Myelination A Practical Pictorial Review. *Neuroimaging Clinics of NA*, 23, 183–195. <https://doi.org/10.1016/j.nic.2012.12.001>
- Breiman, L. (2001). Random Forests. *Machine Learning 2001* 45:1, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Brouwer, M. J., De Vries, L. S., Groenendaal, F., Koopman, C., Pistorius, L. R., Mulder, E. J. H., & Benders, M. J. N. L. (2012). New reference values for the neonatal cerebral ventricles. *Radiology*, 262(1), 224–233. <https://doi.org/10.1148/RADIOL.11110334/ASSET/IMAGES/LARGE/110334T05.JPG>
- Carney, O., Hughes, E., Tusor, N., Dimitrova, R., Arulkumaran, S., Baruteau, K. P., Collado, A. E., Cordero-Grande, L., Chew, A., Falconer, S., Allsop, J. M., Rueckert, D., Hajnal, J., Edwards, A. D., & Rutherford, M. (2021). Incidental findings on brain MR imaging of asymptomatic term neonates in the Developing Human Connectome Project. *EClinicalMedicine*, 38, 100984. <https://doi.org/10.1016/J.ECLINM.2021.100984>
- Chagué, P., Marro, B., Fadili, S., Houot, M., Morin, A., Samper-González, J., Beunon, P., Arrivé, L., Dormont, D., Dubois, B., Teichmann, M., Epelbaum, S., & Colliot, O. (2021). Radiological classification of dementia from anatomical MRI assisted by machine learning-derived maps. *Journal of Neuroradiology*, 48(6), 412–418. <https://doi.org/10.1016/J.NEURAD.2020.04.004>
- Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), 1247–1250. <https://doi.org/10.5194/GMD-7-1247-2014>

- Chang, C. L., Chiu, N. C., Ho, C. S., & Li, S. T. (2006). Frontal horn cysts in normal neonates. *Brain and Development*, 28(7), 426–430. <https://doi.org/10.1016/J.BRAINDEV.2006.01.002>
- Charidimou, A., Boulouis, G., Pasi, M., Auriel, E., Van Etten, E. S., Haley, K., Ayres, A., Schwab, K. M., Martinez-Ramirez, S., Goldstein, J. N., Rosand, J., Viswanathan, A., Greenberg, S. M., & Gurol, M. E. (2017). MRI-visible perivascular spaces in cerebral amyloid angiopathy and hypertensive arteriopathy. *Neurology*, 88(12), 1157. <https://doi.org/10.1212/WNL.00000000000003746>
- Chaturvedi, A., Chaturvedi, A., Stanescu, A. L., Blickman, J. G., & Meyers, S. P. (2018). Mechanical birth-related trauma to the neonate: An imaging perspective. In *Insights into Imaging* (Vol. 9, Issue 1). <https://doi.org/10.1007/s13244-017-0586-x>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Chen, X., & Konukoglu, E. (2018). *Unsupervised Detection of Lesions in Brain MRI using constrained adversarial auto-encoders*. <https://doi.org/10.48550/arxiv.1806.04972>
- Chi, J. G., Dooling, E. C., & Gilles, F. H. (1977). Gyral development of the human brain. *Annals of Neurology*, 1(1), 86–93. <https://doi.org/10.1002/ANA.410010109>
- Cicuendez, M., Castaño-León, A., Ramos, A., Hilario, A., Gómez, P. A., & Lagares, A. (2019). The added prognostic value of magnetic resonance imaging in traumatic brain injury: The importance of traumatic axonal injury when performing ordinal logistic regression. *Journal of Neuroradiology*, 46(5), 299–306. <https://doi.org/10.1016/J.NEURAD.2018.08.001>

Clinical radiology UK workforce census 2020 report. (2021).

Clouchoux, C., Guizard, N., Evans, A. C., Du Plessis, A. J., & Limperopoulos, C. (2012). Normative fetal brain growth by quantitative in vivo magnetic resonance imaging. *American Journal of Obstetrics and Gynecology*, 206(2), 173.e1-173.e8. <https://doi.org/10.1016/J.AJOG.2011.10.002>

Colominas, M. A., Schlotthauer, G., & Torres, M. E. (2014). Improved complete ensemble EMD: A suitable tool for biomedical signal processing. *Biomedical Signal Processing and Control*, 14(1), 19–29. <https://doi.org/10.1016/J.BSPC.2014.06.009>

Cortes, C., Vapnik, V., & Saitta, L. (1995). Support-vector networks. *Machine Learning* 1995 20:3, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>

Counsell, S. J., Rutherford, M. A., Cowan, F. M., & Edwards, A. D. (2003). *Magnetic resonance imaging of preterm brain injury*. <https://doi.org/10.1136/fn.88.4.F269>

Coupé, P., Tourdias, T., Linck, P., Romero, J. E., & Manjón, J. V. (2018). LesionBrain: An online tool for white matter lesion segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11075 LNCS, 95–103. https://doi.org/10.1007/978-3-030-00500-9_11/FIGURES/1

Cowan, F., Rutherford, M., Groenendaal, F., Eken, P., Mercuri, E., Bydder, G. M., Meiners, L. C., Dubowitz, L. M. S., & De Vries, L. S. (2003). Origin and timing of brain lesions in term infants with neonatal encephalopathy. *The Lancet*, 361(9359), 736–742. [https://doi.org/10.1016/S0140-6736\(03\)12658-X](https://doi.org/10.1016/S0140-6736(03)12658-X)

De Coene, B., Hajnal, J. V, Gatehouse, P., Longmore, D. B., White, S. J., Oatridge, A., Pennock, J. M., Young, I. R., & Bydder, G. M. (1992). MR of the brain using fluid-

- attenuated inversion recovery (FLAIR) pulse sequences. *American Journal of Neuroradiology*, 13(6).
- de Graaf-Peters, V. B., & Hadders-Algra, M. (2006). Ontogeny of the human central nervous system: What is happening when? *Early Human Development*, 82(4), 257–266. <https://doi.org/10.1016/J.EARLHUMDEV.2005.10.013>
- De Vries, L. S., Roelants-van Rijn, A. M., Rademaker, K. J., Van Haastert, I. C., Beek, F. J., & Groenendaal, F. (2001). Unilateral parenchymal haemorrhagic infarction in the preterm infant. *European Journal of Paediatric Neurology*, 5(4), 139–149. <https://doi.org/10.1053/EJPN.2001.0494>
- Delannoy, Q., Pham, C. H., Cazorla, C., Tor-Díez, C., Dollé, G., Meunier, H., Bednarek, N., Fablet, R., Passat, N., & Rousseau, F. (2020). SegSRGAN: Super-resolution and segmentation using generative adversarial networks — Application to neonatal brain MRI. *Computers in Biology and Medicine*, 120, 103755. <https://doi.org/10.1016/J.COMPBIOMED.2020.103755>
- Dereymaeker, A., Matic, V., Vervisch, J., Cherian, P. J., Ansari, A. H., De Wel, O., Govaert, P., De Vos, M., Van Huffel, S., Naulaers, G., & Jansen, K. (2019). Automated EEG background analysis to identify neonates with hypoxic-ischemic encephalopathy treated with hypothermia at risk for adverse outcome: A pilot study. *Pediatrics & Neonatology*, 60(1), 50–58. <https://doi.org/10.1016/J.PEDNEO.2018.03.010>
- Dimitriadis, S. I., Liparas, D., & Tsolaki, M. N. (2018). Random forest feature selection, fusion and ensemble strategy: Combining multiple morphological MRI measures to discriminate among healthy elderly, MCI, cMCI and Alzheimer's disease patients: From the Alzheimer's disease neuroimaging initiative (ADNI) database. *Journal of Neuroscience Methods*, 302, 14–23. <https://doi.org/10.1016/J.JNEUMETH.2017.12.010>

- Dimitrova, R., Arulkumaran, S., Carney, O., Chew, A., Falconer, S., Ciarrusta, J., Wolfers, T., Bataille, D., Cordero-Grande, L., Price, A. N., Teixeira, R. P. A. G., Hughes, E., Egloff, A., Hutter, J., Makropoulos, A., Robinson, E. C., Schuh, A., Vecchiato, K., Steinweg, J. K., ... Edwards, A. D. (2021). Phenotyping the Preterm Brain: Characterizing Individual Deviations From Normative Volumetric Development in Two Large Infant Cohorts. *Cerebral Cortex*, 31(8), 3665–3677. <https://doi.org/10.1093/CERCOR/BHAB039>
- Dimitrova, R., Pietsch, M., Ciarrusta, J., Fitzgibbon, S. P., Williams, L. Z. J., Christiaens, D., Cordero-Grande, L., Bataille, D., Makropoulos, A., Schuh, A., Price, A. N., Hutter, J., Teixeira, R. P., Hughes, E., Chew, A., Falconer, S., Carney, O., Egloff, A., Tournier, J. D., ... O’Muircheartaigh, J. (2021). Preterm birth alters the development of cortical microstructure and morphology at term-equivalent age. *NeuroImage*, 243, 118488. <https://doi.org/10.1016/J.NEUROIMAGE.2021.118488>
- Doi, K. (2007). Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31(4–5), 198–211. <https://doi.org/10.1016/J.COMPMEDIMAG.2007.02.002>
- Dorner, R. A., Burton, V. J., Allen, M. C., Robinson, S., & Soares, B. P. (2018). Preterm neuroimaging and neurodevelopmental outcome: a focus on intraventricular hemorrhage, post-hemorrhagic hydrocephalus, and associated brain injury. *Journal of Perinatology* 2018 38:11, 38(11), 1431–1443. <https://doi.org/10.1038/s41372-018-0209-5>
- Doyle, O. M., Ashburner, J., Zelaya, F. O., Williams, S. C. R., Mehta, M. A., & Marquand, A. F. (2013). Multivariate decoding of brain images using ordinal regression. *NeuroImage*, 81, 347–357. <https://doi.org/10.1016/J.NEUROIMAGE.2013.05.036>
- Dyet, L. E., Kennea, N., Counsell, S. J., Maalouf, E. F., Ajayi-Obe, M., Duggan, P. J., Harrison, M., Allsop, J. M., Hajnal, J., Herlihy, A. H., Edwards, B., Laroche, S., Cowan, F. M.,

- Rutherford, M. A., & Edwards, A. D. (2006). Natural History of Brain Lesions in Extremely Preterm Infants Studied With Serial Magnetic Resonance Imaging From Birth and Neurodevelopmental Assessment. *Pediatrics*, 118(2), 536–548. <https://doi.org/10.1542/PEDS.2005-1866>
- Dyrba, M., Grothe, M. J., Mohammadi, A., Binder, H., Kirste, T., & Teipel, S. J. (2018). Comparison of Different Hypotheses Regarding the Spread of Alzheimer’s Disease Using Markov Random Fields and Multimodal Imaging. *Journal of Alzheimer’s Disease*, 65(3), 731–746. <https://doi.org/10.3233/JAD-161197>
- Elster, A. D., & Jonathan H. Burdette. (1994). *Questions & answers in magnetic resonance imaging*.
- Encyclopedia of Machine Learning. (2010). *Encyclopedia of Machine Learning*. <https://doi.org/10.1007/978-0-387-30164-8>
- Essa, E., Aldesouky, D., Hussein, S. E., & Rashad, M. Z. (2020). Neuro-fuzzy patch-wise R-CNN for multiple sclerosis segmentation. *Medical and Biological Engineering and Computing*, 58(9), 2161–2175. <https://doi.org/10.1007/S11517-020-02225-6/TABLES/8>
- Fetit, A. E., Novak, J., Rodriguez, D., Auer, D. P., Clark, C. A., Grundy, R. G., Peet, A. C., & Arvanitis, T. N. (2018). Radiomics in paediatric neuro-oncology: A multicentre study on MRI texture analysis. *NMR in Biomedicine*, 31(1), e3781. <https://doi.org/10.1002/NBM.3781>
- Finck, T., Schinz, D., Grundl, L., Eisawy, R., Yigitsoy, M., Moosbauer, J., Pfister, F., & Wiestler, B. (2021). Automated Pathology Detection and Patient Triage in Routinely Acquired Head Computed Tomography Scans. *Investigative Radiology*, 56(9), 571–578. <https://doi.org/10.1097/RLI.0000000000000775>

- Fineschi, V., Viola, R. V., La Russa, R., Santurro, A., & Frati, P. (2017). A Controversial Medicolegal Issue: Timing the Onset of Perinatal Hypoxic-Ischemic Brain Injury. *Mediators of Inflammation*, 2017. <https://doi.org/10.1155/2017/6024959>
- Folkerth, R. D. (2006). Periventricular leukomalacia: Overview and recent findings. *Pediatric and Developmental Pathology*, 9(1), 3–13. https://doi.org/10.2350/06-01-0024.1/ASSET/IMAGES/LARGE/10.2350_06-01-0024.1-FIG2.JPEG
- Forkert, N. D., Verleger, T., Cheng, B., Thomalla, G., Hilgetag, C. C., & Fiehler, J. (2015). Multiclass Support Vector Machine-Based Lesion Mapping Predicts Functional Outcome in Ischemic Stroke Patients. *PLOS ONE*, 10(6), e0129569. <https://doi.org/10.1371/JOURNAL.PONE.0129569>
- Galloway, M. M. (1975). Texture analysis using gray level run lengths. *Computer Graphics and Image Processing*, 4(2), 172–179. [https://doi.org/10.1016/S0146-664X\(75\)80008-6](https://doi.org/10.1016/S0146-664X(75)80008-6)
- Garcia, K. E., Kroenke, C. D., & Bayly, P. V. (2018). Mechanics of cortical folding: stress, growth and stability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1759). <https://doi.org/10.1098/RSTB.2017.0321>
- Gardner, J. R., Pleiss, G., Bindel, D., Weinberger, K. Q., & Wilson, A. G. (2018). *GPYTORCH: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration*. Retrieved November 15, 2022, from <https://pytorch.ai>.
- Ghosh, N., Sun, Y., Bhanu, B., Ashwal, S., & Obenaus, A. (2014). Automated detection of brain abnormalities in neonatal hypoxia ischemic injury from MR images. *Medical Image Analysis*, 18(7). <https://doi.org/10.1016/j.media.2014.05.002>
- Gilmore, J. H., Schmitt, J. E., Knickmeyer, R. C., Smith, J. K., Lin, W., Styner, M., Gerig, G., & Neale, M. C. (2010). Genetic and environmental contributions to neonatal brain

- structure: A twin study. *Human Brain Mapping*, 31(8), 1174–1182.
<https://doi.org/10.1002/HBM.20926>
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11), 139–144. <https://doi.org/10.1145/3422622>
- Gopagondanahalli, K. R., Li, J., Fahey, M. C., Hunt, R. W., Jenkin, G., Miller, S. L., & Malhotra, A. (2016). Preterm Hypoxic–Ischemic Encephalopathy. *Frontiers in Pediatrics*, 0(OCT), 114. <https://doi.org/10.3389/FPED.2016.00114>
- Granberg, T., Fan, Q., Treaba, C. A., Ouellette, R., Herranz, E., Mangeat, G., Louapre, C., Cohen-Adad, J., Klawiter, E. C., Sloane, J. A., & Mainero, C. (2017). In vivo characterization of cortical and white matter neuroaxonal pathology in early multiple sclerosis. *Brain*, 140(11), 2912–2926. <https://doi.org/10.1093/BRAIN/AWX247>
- Groeschel, S., Vollmer, B., King, M. D., & Connelly, A. (2010). Developmental changes in cerebral grey and white matter volume from infancy to adulthood. *International Journal of Developmental Neuroscience*, 28(6), 481–489.
<https://doi.org/10.1016/j.ijdevneu.2010.06.004>
- Guizard, N., Coupé, P., Fonov, V. S., Manjón, J. V., Arnold, D. L., & Collins, D. L. (2015). Rotation-invariant multi-contrast non-local means for MS lesion segmentation. *NeuroImage: Clinical*, 8, 376–389. <https://doi.org/10.1016/J.NICL.2015.05.001>
- Guo, L., Wang, D., Bo, G., Zhang, H., Tao, W., & Shi, Y. (2016). Early identification of hypoxic-ischemic encephalopathy by combination of magnetic resonance (MR) imaging and proton MR spectroscopy. *Experimental and Therapeutic Medicine*, 12(5), 2835–2842. <https://doi.org/10.3892/ETM.2016.3740/HTML>

- Hampson, M., Driesen, N., Roth, J. K., Gore, J. C., & Constable, R. T. (2010). Functional connectivity between task-positive and task-negative brain areas and its relation to working memory performance. *Magnetic Resonance Imaging*, 28(8), 1051–1057. <https://doi.org/10.1016/J.MRI.2010.03.021>
- Haralick, R. M., Dinstein, I., & Shanmugam, K. (1973). Textural Features for Image Classification. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3(6), 610–621. <https://doi.org/10.1109/TSMC.1973.4309314>
- Hayman, M., van Wezel-Meijler, G., van Straaten, H., Brilstra, E., Groenendaal, F., & de Vries, L. S. (2019). Punctate white-matter lesions in the full-term newborn: Underlying aetiology and outcome. *European Journal of Paediatric Neurology*, 23(2), 280–287. <https://doi.org/10.1016/J.EJPN.2019.01.005>
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- He, L., & Parikh, N. A. (2013). Automated detection of white matter signal abnormality using T2 relaxometry: Application to brain segmentation on term MRI in very preterm infants. *NeuroImage*, 64(1), 328–340. <https://doi.org/10.1016/j.neuroimage.2012.08.081>
- Herrero, M. T., Barcia, C., & Navarro, J. M. (2002). Functional anatomy of thalamus and basal ganglia. *Child's Nervous System* 2002 18:8, 18(8), 386–404. <https://doi.org/10.1007/S00381-002-0604-1>
- Hill, J., Dierker, D., Neil, J., Inder, T., Knutsen, A., Harwell, J., Coalson, T., & Van Essen, D. (2010). *Development/Plasticity/Repair A Surface-Based Analysis of Hemispheric Asymmetries and Folding of Cerebral Cortex in Term-Born Human Infants*. <https://doi.org/10.1523/JNEUROSCI.4682-09.2010>

- Hinojosa-Rodríguez, M., Harmony, T., Carrillo-Prado, C., Van Horn, J. D., Irimia, A., Torgerson, C., & Jacokes, Z. (2017). Clinical neuroimaging in the preterm infant: Diagnosis and prognosis. *NeuroImage: Clinical*, 16, 355–368. <https://doi.org/10.1016/J.NICL.2017.08.015>
- Hinton, G. E., & Zemel, R. S. (1993). *Autoencoders, Minimum Description Length and Helmholtz Free Energy*.
- Holland, D., Chang, L., Ernst, T. M., Curran, M., Buchthal, S. D., Alicata, D., Skranes, J., Johansen, H., Hernandez, A., Yamakawa, R., Kuperman, J. M., & Dale, A. M. (2014). Structural Growth Trajectories and Rates of Change in the First 3 Months of Infant Brain Development. *JAMA Neurology*, 71(10), 1266–1274. <https://doi.org/10.1001/JAMANEUROL.2014.1638>
- Hope, T. M. H., Seghier, M. L., Leff, A. P., & Price, C. J. (2013). Predicting outcome and recovery after stroke with lesions extracted from MRI images. *NeuroImage: Clinical*, 2(1), 424–433. <https://doi.org/10.1016/J.NICL.2013.03.005>
- Hortensius, L. M., Dijkshoorn, A. B. C., Ecury-Goossen, G. M., Steggerda, S. J., Hoebeek, F. E., Benders, M. J. N. L., & Dudink, J. (2018). Neurodevelopmental consequences of preterm isolated cerebellar hemorrhage: A systematic review. *Pediatrics*, 142(5), 20180609. <https://doi.org/10.1542/PEDS.2018-0609/38536>
- Imai, K., De Vries, L. S., Alderliesten, T., Wagenaar, N., Van Der Aa, N. E., Lequin, M. H., Benders, M. J. N. L., Van Haastert, I. C., & Groenendaal, F. (2018). MRI Changes in the Thalamus and Basal Ganglia of Full-Term Neonates with Perinatal Asphyxia. *Neonatology*, 114(3). <https://doi.org/10.1159/000489159>
- Inder, T. E., Wells, S. J., Mogridge, N. B., Spencer, C., & Volpe, J. J. (2003). Defining the nature of the cerebral abnormalities in the premature infant: a qualitative magnetic

resonance imaging study. *The Journal of Pediatrics*, 143(2), 171–179.
[https://doi.org/10.1067/S0022-3476\(03\)00357-3](https://doi.org/10.1067/S0022-3476(03)00357-3)

Introduction to Linear Regression Analysis - Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining - Google Books. (2021). Retrieved January 25, 2023, from
https://books.google.co.uk/books?hl=en&lr=&id=tCIgEAAQBAJ&oi=fnd&pg=PR13&dq=linear+regression&ots=lfybYskZPp&sig=yEUdc7MQFXiNemRoy3Y2opxQmyY&redir_esc=y#v=onepage&q=linear%20regression&f=false

Jarvis, D. A., Finney, C. R., & Griffiths, P. D. (2019). Normative volume measurements of the fetal intra-cranial compartments using 3D volume in utero MR imaging. *European Radiology*, 29(7), 3488–3495. <https://doi.org/10.1007/S00330-018-5938-5/TABLES/2>

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, 17(2), 825–841. <https://doi.org/10.1006/NIMG.2002.1132>

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2), 143–156. [https://doi.org/10.1016/S1361-8415\(01\)00036-6](https://doi.org/10.1016/S1361-8415(01)00036-6)

Jog, A., Carass, A., Pham, D. L., & Prince, J. L. (2015). Multi-Output Decision Trees for Lesion Segmentation in Multiple Sclerosis. *Proceedings of SPIE--the International Society for Optical Engineering*, 9413, 94131C. <https://doi.org/10.1117/12.2082157>

Kaplan, E., Altunisik, E., Ekmekyapar Firat, Y., Datta Barua, P., Dogan, S., Baygin, M., Burak Demir, F., Tuncer, T., Palmer, E., Tan, R. S., Yu, P., Soar, J., Fujita, H., & Rajendra Acharya, U. (2022). Novel nested patch-based feature extraction model for automated Parkinson's Disease symptom classification using MRI images. *Computer Methods and Programs in Biomedicine*, 224, 107030. <https://doi.org/10.1016/J.CMPB.2022.107030>

- Kaur, T., Saini, B. S., & Gupta, S. (2019). An adaptive fuzzy K-nearest neighbor approach for MR brain tumor image classification using parameter free bat optimization algorithm. *Multimedia Tools and Applications*, 78(15), 21853–21890. <https://doi.org/10.1007/S11042-019-7498-3/TABLES/12>
- Kelly, C. J., Arulkumaran, S., Pereira, C. T., Cordero-Grande, L., Hughes, E. J., Pedro, R., Teixeira, A. G., Steinweg, J. K., Victor, S., Pushparajah, K., Hajnal, J. V, Simpson, J., Edwards, A. D., Rutherford, M. A., & Counsell, S. J. (2019). Neuroimaging findings in newborns with congenital heart disease prior to surgery: an observational study What is already known on this topic? *Arch Dis Child*, 104, 1042–1048. <https://doi.org/10.1136/archdischild-2018-314822>
- Kersbergen, K. J., Benders, M. J. N. L., Groenendaal, F., Koopman-Esseboom, C., Nievelstein, R. A. J., Van Haastert, I. C., & De Vries, L. S. (2014). Different Patterns of Punctate White Matter Lesions in Serially Scanned Preterm Infants. *PLOS ONE*, 9(10), e108904. <https://doi.org/10.1371/JOURNAL.PONE.0108904>
- Khawaja, O., & Volpe, J. J. (2007). *Pathogenesis of cerebral white matter injury of prematurity PERIVENTRICULAR LEUKOMALACIA AND ENCEPHALOPATHY OF PREMATUREITY*. <https://doi.org/10.1136/adc.2006.108837>
- Kim, M. J., Loucks, R. A., Palmer, A. L., Brown, A. C., Solomon, K. M., Marchante, A. N., & Whalen, P. J. (2011). The structural and functional connectivity of the amygdala: From normal emotion to pathological anxiety. *Behavioural Brain Research*, 223(2), 403–410. <https://doi.org/10.1016/J.BBR.2011.04.025>
- Kingma, D. P., & Ba, J. L. (2014). Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. <https://doi.org/10.48550/arxiv.1412.6980>

- Kleinbaum, D. G., & Klein, M. (2010). *Logistic Regression*. <https://doi.org/10.1007/978-1-4419-1742-3>
- Kline-Fath, B. M., Horn, P. S., Yuan, W., Merhar, S., Venkatesan, C., Thomas, C. W., & Schapiro, M. B. (2018). Conventional MRI scan and DTI imaging show more severe brain injury in neonates with hypoxic-ischemic encephalopathy and seizures. *Early Human Development*, 122. <https://doi.org/10.1016/j.earlhumdev.2018.05.008>
- Knickmeyer, R. C., Gouttard, S., Kang, C., Evans, D., Wilber, K., Smith, J. K., Hamer, R. M., Lin, W., Gerig, G., & Gilmore, J. H. (2008). *Development/Plasticity/Repair A Structural MRI Study of Human Brain Development from Birth to 2 Years*. <https://doi.org/10.1523/JNEUROSCI.3479-08.2008>
- Knief, H. C., Madesta, F., Schneider, T., Hanning, U., Schönfeld, M. H., Schön, G., Fiehler, J., Gauer, T., Werner, R., & Gellissen, S. (2019). Radiomics of brain MRI: Utility in prediction of metastatic tumor type. *Radiology*, 290(3), 479–487. <https://doi.org/10.1148/RADIOL.2018180946/ASSET/IMAGES/LARGE/RADIOL.2018180946.TBL4.JPEG>
- Kolb, B., & Whishaw, I. Q. (2021). *Fundamentals of Human Neuropsychology* (8th ed.). Worth Publishers.
- Krishnan, P., & Shroff, M. (2016). Neuroimaging in Neonatal Hypoxic Ischemic Encephalopathy. *The Indian Journal of Pediatrics* 2016 83:9, 83(9), 995–1002. <https://doi.org/10.1007/S12098-016-2042-1>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *ImageNet Classification with Deep Convolutional Neural Networks*. Retrieved January 1, 2024, from <http://code.google.com/p/cuda-convnet/>

- Kumar, P., Nagar, P., Arora, C., & Gupta, A. (2018). U-Segnet: Fully Convolutional Neural Network Based Automated Brain Tissue Segmentation Tool. *Proceedings - International Conference on Image Processing, ICIP*, 3503–3507. <https://doi.org/10.1109/ICIP.2018.8451295>
- Kumpulainen, V., Lehtola, S. J., Tuulari, J. J., Silver, E., Copeland, A., Korja, R., Karlsson, H., Karlsson, L., Merisaari, H., Parkkola, R., Saunavaara, J., Lähdesmäki, T., & Scheinin, N. M. (2020). Prevalence and Risk Factors of Incidental Findings in Brain MRIs of Healthy Neonates—The FinnBrain Birth Cohort Study. *Frontiers in Neurology*, 10, 503057. <https://doi.org/10.3389/FNEUR.2019.01347/BIBTEX>
- Kurinczuk, J. J., White-Koning, M., & Badawi, N. (2010). Epidemiology of neonatal encephalopathy and hypoxic–ischaemic encephalopathy. *Early Human Development*, 86(6), 329–338. <https://doi.org/10.1016/J.EARLHUMDEV.2010.05.010>
- Lahmiri, S., & Shmuel, A. (2019). Performance of machine learning methods applied to structural MRI and ADAS cognitive scores in diagnosing Alzheimer’s disease. *Biomedical Signal Processing and Control*, 52, 414–419. <https://doi.org/10.1016/J.BSPC.2018.08.009>
- Lanciego, J. L., Luquin, N., & Obeso, J. A. (2012). Functional Neuroanatomy of the Basal Ganglia. *Cold Spring Harbor Perspectives in Medicine*, 2(12). <https://doi.org/10.1101/CSHPERSPECT.A009621>
- Lekic, T., Klebe, D., Poblete, R., Krafft, P. R., Rolland, W. B., Tang, J., & Zhang, J. H. (2015). Neonatal Brain Hemorrhage (NBH) of Prematurity: Translational Mechanisms of the Vascular-Neural Network. *Current Medicinal Chemistry*, 22(10), 1214. <https://doi.org/10.2174/0929867322666150114152421>

- Lemmon, M. E., Wagner, M. W., Bosemani, T., Carson, K. A., Northington, F. J., Huisman, T. A. G. M., & Poretti, A. (2017). Diffusion Tensor Imaging Detects Occult Cerebellar Injury in Severe Neonatal Hypoxic-Ischemic Encephalopathy. *Developmental Neuroscience*, 39(1–4), 207–214. <https://doi.org/10.1159/000454856>
- Li, A. M., Chau, V., Poskitt, K. J., Sargent, M. A., Lupton, B. A., Hill, A., Roland, E., & Miller, S. P. (2009). White Matter Injury in Term Newborns With Neonatal Encephalopathy. *Pediatric Research* 2009 65:1, 65(1), 85–89. <https://doi.org/10.1203/pdr.0b013e31818912d2>
- Li, G., Wang, L., Shi, F., Lyall, A. E., Lin, W., Gilmore, J. H., & Shen, D. (2014). Mapping Longitudinal Development of Local Cortical Gyrfication in Infants from Birth to 2 Years of Age. *Journal of Neuroscience*, 34(12), 4228–4238. <https://doi.org/10.1523/JNEUROSCI.3976-13.2014>
- Li, H., Parikh, N. A., Wang, J., Merhar, S., Chen, M., Parikh, M., Holland, S., & He, L. (2019). Objective and Automated Detection of Diffuse White Matter Abnormality in Preterm Infants Using Deep Convolutional Neural Networks. *Frontiers in Neuroscience*, 0(JUN), 610. <https://doi.org/10.3389/FNINS.2019.00610>
- Li, S. Z. (2001). Markov Random Field Modeling in Image Analysis. *Computer Science Workbench*. <https://doi.org/10.1007/978-4-431-67044-5>
- Liu, L., Chen, S., Zhang, F., Wu, F. X., Pan, Y., & Wang, J. (2020). Deep convolutional neural network for automatically segmenting acute ischemic stroke lesion in multi-modality MRI. *Neural Computing and Applications*, 32(11), 6545–6558. <https://doi.org/10.1007/S00521-019-04096-X/TABLES/5>
- Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J.,

- Bachmann, L. M., Keane, P. A., & Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, 1(6), e271–e297. [https://doi.org/10.1016/S2589-7500\(19\)30123-2](https://doi.org/10.1016/S2589-7500(19)30123-2)
- Luo, J., Luo, Y., Zeng, H., Reis, C., & Chen, S. (2019). Research Advances of Germinal Matrix Hemorrhage: An Update Review. *Cellular and Molecular Neurobiology*, 39(1), 1–10. <https://doi.org/10.1007/S10571-018-0630-5/FIGURES/4>
- Lynch, J. K., Hirtz, D. G., DeVeber, G., & Nelson, K. B. (2002). Report of the National Institute of Neurological Disorders and Stroke Workshop on Perinatal and Childhood Stroke. *Pediatrics*, 109(1), 116–123. <https://doi.org/10.1542/PEDS.109.1.116>
- MacKay, M. T., Slavova, N., Pastore-Wapp, M., Grunt, S., Stojanovski, B., Donath, S., & Steinlin, M. (2020). Pediatric ASPECTS predicts outcomes following acute symptomatic neonatal arterial stroke. *Neurology*, 94(12), e1259–e1270. <https://doi.org/10.1212/WNL.00000000000009136>
- Makropoulos, A., Robinson, E. C., Schuh, A., Wright, R., Fitzgibbon, S., Bozek, J., Counsell, S. J., Steinweg, J., Vecchiato, K., Passerat-Palmbach, J., Lenz, G., Mortari, F., Tenev, T., Duff, E. P., Bastiani, M., Cordero-Grande, L., Hughes, E., Tusor, N., Tournier, J. D., ... Rueckert, D. (2018). The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. *NeuroImage*, 173, 88–112. <https://doi.org/10.1016/J.NEUROIMAGE.2018.01.054>
- Mannor, S., Jin, X., Han, J., Jin, X., Han, J., Jin, X., Han, J., & Zhang, X. (2011). K-Means Clustering. *Encyclopedia of Machine Learning*, 563–564. https://doi.org/10.1007/978-0-387-30164-8_425

- Martinez-Biarge, M., Diez-Sebastian, J., Kapellou, O., Gindner, D., Allsop, J. M., Rutherford, M. A., & Cowan, F. M. (2011). Predicting motor outcome and death in term hypoxic-ischemic encephalopathy. *Neurology*, 76(24), 2055–2061. <https://doi.org/10.1212/WNL.0B013E31821F442D>
- Martinez-Biarge, M., Diez-Sebastian, J., Rutherford, M. A., & Cowan, F. M. (2010). Outcomes after central grey matter injury in term perinatal hypoxic-ischaemic encephalopathy. *Early Human Development*, 86(11), 675–682. <https://doi.org/10.1016/J.EARLHUMDEV.2010.08.013>
- McAllister, J. P. (2012). Pathophysiology of congenital and neonatal hydrocephalus. *Seminars in Fetal and Neonatal Medicine*, 17(5), 285–294. <https://doi.org/10.1016/J.SINY.2012.06.004>
- McArdle, C. B., Richardson, C. J., Nicholas, D. A., Mirfakhraee, M., Hayden, C. K., & Amparo, E. G. (1987). Developmental features of the neonatal brain: MR imaging. Part I. Gray-white matter differentiation and myelination. *https://Doi.Org/10.1148/Radiology.162.1.3786767*, 162(1), 223–229. <https://doi.org/10.1148/RADIOLOGY.162.1.3786767>
- Medsker, L., Jain, D. L., & Raton London New York Washington, B. (2001). *RECURRENT NEURAL NETWORKS E dited by Design and Applications*.
- Meng, Y., Li, G., Wang, L., Lin, W., Gilmore, J. H., & Shen, D. (2018). Discovering cortical sulcal folding patterns in neonates using large-scale dataset. *Human Brain Mapping*, 39(9), 3625. <https://doi.org/10.1002/HBM.24199>
- Mercadante, A. A., & Tadi, P. (2022). Neuroanatomy, Gray Matter. *StatPearls*. <https://www.ncbi.nlm.nih.gov/books/NBK553239/>

- Merzoug, A., Benamrane, N., & Taleb-Ahmed, A. (1 C.E.). Lesions Detection of Multiple Sclerosis in 3D Brian MR Images by Using Artificial Immune Systems and Support Vector Machines. <https://Services.Igi-Global.Com/Resolvedoi/Resolve.Asp?Doi=10.4018/978-1-6684-7544-7.Ch033>, 671–685. <https://doi.org/10.4018/978-1-6684-7544-7.CH033>
- Millar, L. J., Shi, L., Hoerder-Suabedissen, A., & Molnár, Z. (2017). Neonatal Hypoxia Ischaemia: Mechanisms, Models, and Therapeutic Challenges. *Frontiers in Cellular Neuroscience*, 0, 78. <https://doi.org/10.3389/FNCEL.2017.00078>
- Miller, J. H., Bardo, D. M. E., & Cornejo, P. (2020). Neonatal Neuroimaging. *Seminars in Pediatric Neurology*, 33. <https://doi.org/10.1016/j.spen.2020.100796>
- Miller, V. (2000). Neonatal cerebral infarction. *Seminars in Pediatric Neurology*, 7(4), 278–288. <https://doi.org/10.1053/SPEN.2000.20076>
- Moore, P. J., Lyons, T. J., & Gallacher, J. (2019). Random forest prediction of Alzheimer’s disease using pairwise selection from time series data. *PLOS ONE*, 14(2), e0211558. <https://doi.org/10.1371/JOURNAL.PONE.0211558>
- Mugler, J. P., & Brookeman, J. R. (1990). Three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP RAGE). *Magnetic Resonance in Medicine*, 15(1), 152–157. <https://doi.org/10.1002/MRM.1910150117>
- Murphy, K., van der Aa, N. E., Negro, S., Groenendaal, F., de Vries, L. S., Viergever, M. A., Boylan, G. B., Benders, M. J. N. L., & Išgum, I. (2017). Automatic quantification of ischemic injury on diffusion-weighted MRI of neonatal hypoxic ischemic encephalopathy. *NeuroImage: Clinical*, 14. <https://doi.org/10.1016/j.nicl.2017.01.005>

- Nazari-Farsani, S., Nyman, M., Karjalainen, T., Bucci, M., Isojärvi, J., & Nummenmaa, L. (2020). Automated segmentation of acute stroke lesions using a data-driven anomaly detection on diffusion weighted MRI. *Journal of Neuroscience Methods*, 333, 108575. <https://doi.org/10.1016/J.JNEUMETH.2019.108575>
- Nishikuni, K., & Ribas, G. C. (2013). Study of fetal and postnatal morphological development of the brain sulci: Laboratory investigation. *Journal of Neurosurgery: Pediatrics*, 11(1), 1–11. <https://doi.org/10.3171/2012.9.PEDS12122>
- O’Muircheartaigh, J., Dean, D. C., Ginestet, C. E., Walker, L., Waskiewicz, N., Lehman, K., Dirks, H., Piryatinsky, I., & Deoni, S. C. L. (2014). White matter development and early cognition in babies and toddlers. *Human Brain Mapping*, 35(9), 4475–4487. <https://doi.org/10.1002/HBM.22488>
- O’Muircheartaigh, J., Robinson, E. C., Pietsch, M., Wolfers, T., Aljabar, P., Grande, L. C., Teixeira, R. P. A. G., Bozek, J., Schuh, A., Makropoulos, A., Batalle, D., Hutter, J., Vecchiato, K., Steinweg, J. K., Fitzgibbon, S., Hughes, E., Price, A. N., Marquand, A., Reuckert, D., ... Edwards, A. D. (2020). Modelling brain development to detect white matter injury in term and preterm born neonates. *Brain*, 143(2), 467–479. <https://doi.org/10.1093/BRAIN/AWZ412>
- Ouwehand, S., Smidt, L. C. A., Dudink, J., Benders, M. J. N. L., De Vries, L. S., Groenendaal, F., & Van Der Aa, N. E. (2020). Predictors of Outcomes in Hypoxic-Ischemic Encephalopathy following Hypothermia: A Meta-Analysis. *Neonatology*, 117(4), 411–427. <https://doi.org/10.1159/000505519>
- Pasternak, J. F., Predey, T. A., & Mikhael, M. A. (1991). Neonatal asphyxia: Vulnerability of basal ganglia, thalamus, and brainstem. *Pediatric Neurology*, 7(2), 147–149. [https://doi.org/10.1016/0887-8994\(91\)90014-C](https://doi.org/10.1016/0887-8994(91)90014-C)

- Pereira, S., Pinto, A., Oliveira, J., Mendrik, A. M., Correia, J. H., & Silva, C. A. (2016). Automatic brain tissue segmentation in MR images using Random Forests and Conditional Random Fields. *Journal of Neuroscience Methods*, 270, 111–123. <https://doi.org/10.1016/J.JNEUMETH.2016.06.017>
- Perez, A., Ritter, S., Brotschi, B., Werner, H., Caflisch, J., Martin, E., & Latal, B. (2013). Long-Term Neurodevelopmental Outcome with Hypoxic-Ischemic Encephalopathy. *The Journal of Pediatrics*, 163(2), 454-459.e1. <https://doi.org/10.1016/J.JPEDS.2013.02.003>
- Perlman, M., & Shah, P. S. (2011). Hypoxic-Ischemic Encephalopathy: Challenges in Outcome and Prediction. *The Journal of Pediatrics*, 158(2), e51–e54. <https://doi.org/10.1016/J.JPEDS.2010.11.014>
- Peterson, B. S., Vohr, B., Staib, L. H., Cannistraci, C. J., Dolberg, A., Schneider, K. C., Katz, K. H., Westerveld, M., Sparrow, S., Anderson, A. W., Duncan, C. C., Makuch, R. W., Gore, J. C., & Ment, L. R. (2000). Regional Brain Volume Abnormalities and Long-term Cognitive Outcome in Preterm Infants. *JAMA*, 284(15), 1939–1947. <https://doi.org/10.1001/JAMA.284.15.1939>
- Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, 4(2), 1883. <https://doi.org/10.4249/SCHOLARPEDIA.1883>
- Piñeiro-Ramos, J. D., Núñez-Ramiro, A., Llorens-Salvador, R., Parra-Llorca, A., Sánchez-Illana, Á., Quintás, G., Boronat-González, N., Martínez-Rodilla, J., Kuligowski, J., & Vento, M. (2020). Metabolic Phenotypes of Hypoxic-Ischemic Encephalopathy with Normal vs. Pathologic Magnetic Resonance Imaging Outcomes. *Metabolites* 2020, Vol. 10, Page 109, 10(3), 109. <https://doi.org/10.3390/METABO10030109>
- Pinto, M. F., Oliveira, H., Batista, S., Cruz, L., Pinto, M., Correia, I., Martins, P., & Teixeira, C. (2020). Prediction of disease progression and outcomes in multiple sclerosis with

- machine learning. *Scientific Reports* 2020 10:1, 10(1), 1–13.
<https://doi.org/10.1038/s41598-020-78212-6>
- Pradilla, G., & Jallo, G. (2007). Arachnoid cysts: case series and review of the literature. *Neurosurgical Focus*, 22(2), 1–4. <https://doi.org/10.3171/FOC.2007.22.2.7>
- Prempunpong, C., Chalak, L. F., Garfinkle, J., Shah, B., Kalra, V., Rollins, N., Boyle, R., Nguyen, K. A., Mir, I., Pappas, A., Montaldo, P., Thayyil, S., Sánchez, P. J., Shankaran, S., Laptook, A. R., & Sant’anna, G. (2017). Prospective research on infants with mild encephalopathy: the PRIME study. *Journal of Perinatology* 2018 38:1, 38(1), 80–85.
<https://doi.org/10.1038/jp.2017.164>
- Purves, D., Augustine, G. J., Fitzpatrick, D., Katz, L. C., LaMantia, A.-S., McNamara, J. O., & Williams, S. M. (2001). *The Blood Supply of the Brain and Spinal Cord*.
<https://www.ncbi.nlm.nih.gov/books/NBK11042/>
- Quattrocchi, C. C., Fariello, G., & Longo, D. (2016). Brainstem tegmental lesions in neonates with hypoxic-ischemic encephalopathy: Magnetic resonance diagnosis and clinical outcome. *World Journal of Radiology*, 8(2), 117. <https://doi.org/10.4329/WJR.V8.I2.117>
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning* 1986 1:1, 1(1), 81–106.
<https://doi.org/10.1007/BF00116251>
- Rajagopalan, V., Scott, J., Habas, P. A., Kim, K., Corbett-Detig, J., Rousseau, F., Barkovich, A. J., Glenn, O. A., & Studholme, C. (2011). *Development/Plasticity/Repair Local Tissue Growth Patterns Underlying Normal Fetal Human Brain Gyriification Quantified In Utero*. <https://doi.org/10.1523/JNEUROSCI.5458-10.2011>
- Rallabandi, V. P. S., Tulpule, K., & Gattu, M. (2020). Automatic classification of cognitively normal, mild cognitive impairment and Alzheimer’s disease using structural MRI

- analysis. *Informatics in Medicine Unlocked*, 18, 100305.
<https://doi.org/10.1016/J.IMU.2020.100305>
- Ramenghi, L. A., Fumagalli, M., Righini, A., Bassi, L., Groppo, M., Parazzini, C., Bianchini, E., Triulzi, F., & Mosca, F. (2007). Magnetic resonance imaging assessment of brain maturation in preterm neonates with punctate white matter lesions. *Neuroradiology*, 49(2), 161–167. <https://doi.org/10.1007/S00234-006-0176-Y/FIGURES/5>
- Ramírez, J., Górriz, J. M., Ortiz, A., Martínez-Murcia, F. J., Segovia, F., Salas-Gonzalez, D., Castillo-Barnes, D., Illán, I. A., & Puntonet, C. G. (2018). Ensemble of random forests One vs. Rest classifiers for MCI and AD prediction using ANOVA cortical and subcortical feature selection and partial least squares. *Journal of Neuroscience Methods*, 302, 47–57. <https://doi.org/10.1016/J.JNEUMETH.2017.12.005>
- Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Retrieved July 8, 2019, from www.GaussianProcess.org/gpml
- Reid, S. M., Dagia, C. D., Ditchfield, M. R., & Reddihough, D. S. (2015). Grey matter injury patterns in cerebral palsy: associations between structural involvement on MRI and clinical outcomes. *Developmental Medicine & Child Neurology*, 57(12), 1159–1167. <https://doi.org/10.1111/DMCN.12800>
- Rezaei, M., Yang, H., & Meinel, C. (2017). *Brain Abnormality Detection by Deep Convolutional Neural Network*. <https://doi.org/10.48550/arxiv.1708.05206>
- Richhariya, B., Tanveer, M., & Rashid, A. H. (2020). Diagnosis of Alzheimer's disease using universum support vector machine based recursive feature elimination (USVM-RFE). *Biomedical Signal Processing and Control*, 59, 101903. <https://doi.org/10.1016/J.BSPC.2020.101903>

- Rollins, N., Booth, T., Morriss, M. C., Sanchez, P., Heyne, R., & Chalak, L. (2014). Predictive Value of Neonatal MRI Showing No or Minor Degrees of Brain Injury after Hypothermia. *Pediatric Neurology*, 50(5), 447. <https://doi.org/10.1016/J.PEDIATRNEUROL.2014.01.013>
- Ronan, L., & Fletcher, P. C. (2015). From genes to folds: a review of cortical gyrification theory. *Brain Structure and Function*, 220(5), 2475–2483. <https://doi.org/10.1007/S00429-014-0961-Z/FIGURES/2>
- Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Retrieved January 1, 2024, from <http://lmb.informatik.uni-freiburg.de/>
- Rosipal, R., & Krämer, N. (2006). Overview and Recent Advances in Partial Least Squares. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3940 LNCS, 34–51. https://doi.org/10.1007/11752790_2
- Roy, P. K., Bhuiyan, A., Janke, A., Desmond, P. M., Wong, T. Y., Abhayaratna, W. P., Storey, E., & Ramamohanarao, K. (2015). Automatic white matter lesion segmentation using contrast enhanced FLAIR intensity and Markov Random Field. *Computerized Medical Imaging and Graphics*, 45, 102–111. <https://doi.org/10.1016/J.COMPMEDIMAG.2015.08.005>
- Roy, S., Butman, J. A., Reich, D. S., Calabresi, P. A., & Pham, D. L. (2018). *Multiple Sclerosis Lesion Segmentation from Brain MRI via Fully Convolutional Neural Networks*. <https://doi.org/10.48550/arxiv.1803.09172>

- Runchey, S., & McGee, S. (2010). Does This Patient Have a Hemorrhagic Stroke?: Clinical Findings Distinguishing Hemorrhagic Stroke From Ischemic Stroke. *JAMA*, 303(22), 2280–2286. <https://doi.org/10.1001/JAMA.2010.754>
- Rutherford, M. A. (2002). *MRI of the neonatal brain*. 375.
- Rutherford, M. A., Ramenghi, L. A., & Cowan, F. M. (2012). Neonatal stroke. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 97(5), F377–F384. <https://doi.org/10.1136/FETALNEONATAL-2010-196451>
- Rutherford, M., Biarge, M. M., Allsop, J., Counsell, S., & Cowan, F. (2010). MRI of perinatal brain injury. *Pediatric Radiology*, 40(6), 819–833. <https://doi.org/10.1007/S00247-010-1620-Z/FIGURES/17>
- Rutherford, M., Malamateniou, C., McGuinness, A., Allsop, J., Biarge, M. M., & Counsell, S. (2010). Magnetic resonance imaging in hypoxic-ischaemic encephalopathy. *Early Human Development*, 86(6), 351–360. <https://doi.org/10.1016/J.EARLHUMDEV.2010.05.014>
- Rutherford, M., Malamateniou, C., Zeka, J., & Counsell, S. (2004). MR imaging of the neonatal brain at 3 Tesla. *European Journal of Paediatric Neurology*, 8(6), 281–289. <https://doi.org/10.1016/J.EJPN.2004.08.003>
- Saccà, V., Sarica, A., Novellino, F., Barone, S., Tallarico, T., Filippelli, E., Granata, A., Chiriaco, C., Bruno Bossio, R., Valentino, P., & Quattrone, A. (2019). Evaluation of machine learning algorithms performance for the prediction of early multiple sclerosis from resting-state fMRI connectivity data. *Brain Imaging and Behavior*, 13(4), 1103–1114. <https://doi.org/10.1007/S11682-018-9926-9/TABLES/3>

- Saladi, S., & Amutha Prabha, N. (2018). MRI brain segmentation in combination of clustering methods with Markov random field. *International Journal of Imaging Systems and Technology*, 28(3), 207–216. <https://doi.org/10.1002/IMA.22271>
- Salvador, R., Radua, J., Canales-Rodríguez, E. J., Solanes, A., Sarroa, S., Goikolea, J. M., Valiente, A., Montea, G. C., Natividad, M. D. C., Guerrero-Pedraza, A., Moro, N., Fernández-Corcuera, P., Amann, B. L., Maristany, T., Vieta, E., McKenna, P. J., & Pomarol-Clote, E. (2017). Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis. *PLOS ONE*, 12(4), e0175683. <https://doi.org/10.1371/JOURNAL.PONE.0175683>
- Sargent, M. A., Poskitt, K. J., Roland, E. H., Hill, A., & Henderson, G. (2004). *Cerebellar Vermian Atrophy after Neonatal Hypoxic-Ischemic Encephalopathy*.
- Schiavi, S., Azzari, A., Mensi, A., Graziano, N., Daducci, A., Bicego, M., Inglese, M., & Petracca, M. (2022). Classification of multiple sclerosis patients based on structural disconnection: A robust feature selection approach. *Journal of Neuroimaging*. <https://doi.org/10.1111/JON.12991>
- Schiffrin, B. S., & Ater, S. (2006). Fetal hypoxic and ischemic injuries. *Current Opinion in Obstetrics and Gynecology*, 18(2), 112–122. <https://doi.org/10.1097/01.GCO.0000192984.15095.7C>
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85–117. <https://doi.org/10.1016/J.NEUNET.2014.09.003>
- Schuh, A., Makropoulos, A., Robinson, E. C., Cordero-Grande, L., Hughes, E., Hutter, J., Price, A. N., Murgasova, M., Pedro G Teixeira, R. A., Tusor, N., Steinweg, J. K., Victor, S., Rutherford, M. A., Hajnal, J. V, David Edwards, A., & Rueckert, D. (2018). *Unbiased*

construction of a temporally consistent morphological atlas of neonatal brain development. <https://doi.org/10.1101/251512>

Shankaran, S., Laptook, A. R., Ehrenkranz, R. A., Tyson, J. E., McDonald, S. A., Donovan, E. F., Fanaroff, A. A., Poole, W. K., Wright, L. L., Higgins, R. D., Finer, N. N., Carlo, W. A., Duara, S., Oh, W., Cotten, C. M., Stevenson, D. K., Stoll, B. J., Lemons, J. A., Guillet, R., & Jobe, A. H. (2005). Whole-Body Hypothermia for Neonates with Hypoxic–Ischemic Encephalopathy. *Https://Doi.Org/10.1056/NEJMcps050929*, 353(15), 1574–1584. <https://doi.org/10.1056/NEJMCPS050929>

Shankaran, S., Laptook, A. R., McDonald, S. A., Hintz, S. R., Barnes, P. D., Das, A., & Higgins, R. D. (2017). Acute Perinatal Sentinel Events, Neonatal Brain Injury Pattern, and Outcome of Infants Undergoing a Trial of Hypothermia for Neonatal Hypoxic-Ischemic Encephalopathy. *The Journal of Pediatrics*, 180, 275-278.e2. <https://doi.org/10.1016/J.JPEDS.2016.09.026>

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, 17(3), 143–155. <https://doi.org/10.1002/HBM.10062>

Sørensen, L., & Nielsen, M. (2018). Ensemble support vector machine classification of dementia using structural MRI and mini-mental state examination. *Journal of Neuroscience Methods*, 302, 66–74. <https://doi.org/10.1016/J.JNEUMETH.2018.01.003>

Sprung-Much, T., & Petrides, M. (2018). Morphological patterns and spatial probability maps of two defining sulci of the posterior ventrolateral frontal cortex of the human brain: the sulcus diagonalis and the anterior ascending ramus of the lateral fissure. *Brain Structure and Function*, 223(9), 4125–4152. <https://doi.org/10.1007/S00429-018-1733-Y/FIGURES/18>

- Srhoj-Egekher, V., Benders, M. J. N. L., Kersbergen, K. J., Viergever, M. A., & Išgum, I. (2012). *Automatic segmentation of neonatal brain MRI using atlas based segmentation and machine learning approach*. Retrieved November 19, 2018, from <http://neobrain12.isi.uu.nl>
- Steenwijk, M. D., Pouwels, P. J. W., Daams, M., Van Dalen, J. W., Caan, M. W. A., Richard, E., Barkhof, F., & Vrenken, H. (2013). Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). *NeuroImage: Clinical*, 3, 462–469. <https://doi.org/10.1016/J.NICL.2013.10.003>
- Stevenson, N. J., Korotchikova, I., Temko, A., Lightbody, G., Marnane, W. P., & Boylan, G. B. (2013). An automated system for grading eeg abnormality in term neonates with hypoxic-ischaemic encephalopathy. *Annals of Biomedical Engineering*, 41(4), 775–785. <https://doi.org/10.1007/S10439-012-0710-5/FIGURES/6>
- Subbanna, N. K., Rajashekar, D., Cheng, B., Thomalla, G., Fiehler, J., Arbel, T., & Forkert, N. D. (2019). Stroke lesion segmentation in FLAIR MRI datasets using customized Markov random fields. *Frontiers in Neurology*, 10(MAY), 541. <https://doi.org/10.3389/FNEUR.2019.00541/BIBTEX>
- Subbanna, N., Precup, D., Arnold, D., & Arbel, T. (2015). IMaGe: Iterative multilevel probabilistic graphical model for detection and segmentation of multiple sclerosis lesions in brain MRI. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9123, 514–526. https://doi.org/10.1007/978-3-319-19992-4_40/FIGURES/6
- Subudhi, A., Dash, M., & Sabut, S. (2020). Automated segmentation and classification of brain stroke using expectation-maximization and random forest classifier. *Biocybernetics and Biomedical Engineering*, 40(1), 277–289. <https://doi.org/10.1016/J.BBE.2019.04.004>

- Sweeney, E. M., Nguyen, T. D., Kuceyeski, A., Ryan, S. M., Zhang, S., Zexter, L., Wang, Y., & Gauthier, S. A. (2021). Estimation of Multiple Sclerosis lesion age on magnetic resonance imaging. *NeuroImage*, 225, 117451. <https://doi.org/10.1016/J.NEUROIMAGE.2020.117451>
- Szabó, Z., Kapás, Z., Lefkovits, L., Gy", A., Szilágyi, S. M., & Szilágyi, L. (2018). *Automatic segmentation of low-grade brain tumor using a random forest classifier and Gabor features*.
- Tang, Z., Mahmoodi, S., Dasmahapatra, S., Darekar, A., & Vollmer, B. (2020). Ridge Detection and Analysis of Susceptibility-Weighted Magnetic Resonance Imaging in Neonatal Hypoxic-Ischaemic Encephalopathy. *Communications in Computer and Information Science*, 1248 CCIS, 307–318. https://doi.org/10.1007/978-3-030-52791-4_24/TABLES/3
- Tarui, T., Im, K., Madan, N., Madankumar, R., Skotko, B. G., Schwartz, A., Sharr, C., Ralston, S. J., Kitano, R., Akiyama, S., Yun, H. J., Grant, E., & Bianchi, D. W. (2020). Quantitative MRI Analyses of Regional Brain Growth in Living Fetuses with Down Syndrome. *Cerebral Cortex*, 30(1), 382–390. <https://doi.org/10.1093/CERCOR/BHZ094>
- Tencer, L., Reznakova, M., & Cheriet, M. (2012). A new framework for online sketch-based image retrieval in web environment. *2012 11th International Conference on Information Science, Signal Processing and Their Applications, ISSPA 2012*, 1430–1431. <https://doi.org/10.1109/ISSPA.2012.6310525>
- Tribolet, N. (2010). A. R. Crossman, D. Neary, Neuroanatomy. An illustrated colour text. *Acta Neurochirurgica*, 152(11), 2009–2009. <https://doi.org/10.1007/S00701-010-0712-8>
- Trivedi, S. B., Vesoulis, Z. A., Rao, R., Liao, S. M., Shimony, J. S., McKinstry, R. C., & Mathur, A. M. (2017). A validated clinical MRI injury scoring system in neonatal

- hypoxic-ischemic encephalopathy. *Pediatric Radiology*, 47(11), 1491–1499.
<https://doi.org/10.1007/S00247-017-3893-Y/TABLES/5>
- Tusor, N., Benders, M. J., Counsell, S. J., Nongena, P., Ederies, M. A., Falconer, S., Chew, A., Gonzalez-Cinca, N., Hajnal, J. V., Gangadharan, S., Chatzi, V., Kersbergen, K. J., Kennea, N., Azzopardi, D. V., & Edwards, A. D. (2017). Punctate White Matter Lesions Associated With Altered Brain Development And Adverse Motor Outcome In Preterm Infants. *Scientific Reports* 2017 7:1, 7(1), 1–9. <https://doi.org/10.1038/s41598-017-13753-x>
- Vamvakas, A., Williams, S. C., Theodorou, K., Kapsalaki, E., Fountas, K., Kappas, C., Vassiou, K., & Tsougos, I. (2019). Imaging biomarker analysis of advanced multiparametric MRI for glioma grading. *Physica Medica*, 60, 188–198.
<https://doi.org/10.1016/J.EJMP.2019.03.014>
- Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, 180, 68–77. <https://doi.org/10.1016/J.NEUROIMAGE.2017.06.061>
- Vasung, L., Abaci Turk, E., Ferradal, S. L., Sutin, J., Stout, J. N., Ahtam, B., Lin, P. Y., & Grant, P. E. (2019). Exploring early human brain development with structural and physiological neuroimaging. *NeuroImage*, 187, 226.
<https://doi.org/10.1016/J.NEUROIMAGE.2018.07.041>
- Velazquez, M., Lee, Y., & Initiative, for the A. D. N. (2021). Random forest model for feature-based Alzheimer’s disease conversion prediction from early mild cognitive impairment subjects. *PLOS ONE*, 16(4), e0244773.
<https://doi.org/10.1371/JOURNAL.PONE.0244773>

- Volpe, J. J. (2009). Brain injury in premature infants: a complex amalgam of destructive and developmental disturbances. *The Lancet Neurology*, 8(1), 110–124. [https://doi.org/10.1016/S1474-4422\(08\)70294-1](https://doi.org/10.1016/S1474-4422(08)70294-1)
- Volpe, J. J. (2012). Neonatal encephalopathy: An inadequate term for hypoxic–ischemic encephalopathy. *Annals of Neurology*, 72(2), 156–166. <https://doi.org/10.1002/ANA.23647>
- Volpe, J. J., & Zipurksy, A. (2001). Neurobiology of Periventricular Leukomalacia in the Premature Infant. *Pediatric Research* 2001 50:5, 50(5), 553–562. <https://doi.org/10.1203/00006450-200111000-00003>
- Weiss, R. J., Bates, S. V., Song, Y., Zhang, Y., Herzberg, E. M., Chen, Y. C., Gong, M., Chien, I., Zhang, L., Murphy, S. N., Gollub, R. L., Grant, P. E., & Ou, Y. (2019). Mining multi-site clinical data to develop machine learning MRI biomarkers: application to neonatal hypoxic ischemic encephalopathy. *Journal of Translational Medicine* 2019 17:1, 17(1), 1–16. <https://doi.org/10.1186/S12967-019-2119-5>
- Welker, K. M., & Patton, A. (2012). *Assessment of Normal Myelination with Magnetic Resonance Imaging*. <https://doi.org/10.1055/s-0032-1306382>
- Whitworth, M., Bricker, L., & Mullan, C. (2015). Ultrasound for fetal assessment in early pregnancy. *Cochrane Database of Systematic Reviews*, 2015(7). https://doi.org/10.1002/14651858.CD007058.PUB3/MEDIA/CDSR/CD007058/IMAGE_N/NCD007058-CMP-001-38.PNG
- Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1), 79–82. <https://doi.org/10.3354/CR030079>

- Wolf, R. L., Zimmerman, R. A., Clancy, R., & Haselgrove, J. H. (2001). Quantitative apparent diffusion coefficient measurements in term neonates for early detection of hypoxic-ischemic brain injury: Initial experience. *Radiology*, 218(3), 825–833. <https://doi.org/10.1148/RADIOLOGY.218.3.R01FE47825/ASSET/IMAGES/LARGE/R01FE47G4D.JPEG>
- Xin, W., & Chan, J. R. (2020). Myelin plasticity: sculpting circuits in learning and memory. *Nature Reviews Neuroscience* 2020 21:12, 21(12), 682–694. <https://doi.org/10.1038/s41583-020-00379-8>
- Yanase, J., & Triantaphyllou, E. (2019). A systematic survey of computer-aided diagnosis in medicine: Past and present developments. *Expert Systems with Applications*, 138, 112821. <https://doi.org/10.1016/J.ESWA.2019.112821>
- Yang, X.-S. S. (2010). A New Metaheuristic Bat-Inspired Algorithm BT - Nature Inspired Cooperative Strategies for Optimization (NICSO 2010). *Studies in Computational Intelligence*, 284, 65–74. <http://arxiv.org/abs/1004.4170>
- Yuh, E. L., Mukherjee, P., Lingsma, H. F., Yue, J. K., Ferguson, A. R., Gordon, W. A., Valadka, A. B., Schnyer, D. M., Okonkwo, D. O., Maas, A. I. R., & Manley, G. T. (2013). Magnetic resonance imaging improves 3-month outcome prediction in mild traumatic brain injury. *Annals of Neurology*, 73(2), 224–235. <https://doi.org/10.1002/ANA.23783>
- Zaitsev, M., Maclaren, J., & Herbst, M. (2015). Motion Artefacts in MRI: a Complex Problem with Many Partial Solutions. *Journal of Magnetic Resonance Imaging : JMRI*, 42(4), 887. <https://doi.org/10.1002/JMRI.24850>
- Ziegler, G., Ridgway, G. R., Dahnke, R., & Gaser, C. (2014). Individualized Gaussian process-based prediction and detection of local and global gray matter abnormalities in elderly subjects. *NeuroImage*, 97, 333–348. <https://doi.org/10.1016/j.neuroimage.2014.04.018>

Zilles, K., Schleicher, A., Langemann, C., Amunts, K., Morosan, P., Palomero-Gallagher, N., Schormann, T., Mohlberg, H., Bü Rgel, U., Steinmetz, H., Schlaug, G., & Roland, P. E. (1997). Quantitative Analysis of Sulci in the Human Cerebral Cortex: Development, Regional Heterogeneity, Gender Difference, Asymmetry, Intersubject Variability and Cortical Architecture. *Hum. Brain Mapping*, 5, 218–221.
[https://doi.org/10.1002/\(SICI\)1097-0193\(1997\)5:4](https://doi.org/10.1002/(SICI)1097-0193(1997)5:4)

Zurita, M., Montalba, C., Labbé, T., Cruz, J. P., Dalboni da Rocha, J., Tejos, C., Ciampi, E., Cárcamo, C., Sitaram, R., & Uribe, S. (2018). Characterization of relapsing-remitting multiple sclerosis patients using support vector machine classifications of functional and diffusion MRI data. *NeuroImage: Clinical*, 20, 724–730.
<https://doi.org/10.1016/J.NICL.2018.09.002>

Chapter 8. Appendices

8.1 Chapter 3 Publication - Automatic Detection of Neonatal Brain Injury on MRI

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Automatic Detection of Neonatal Brain Injury on MRI." *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis: First International Workshop, ASMUS 2020, and 5th International Workshop, PIPPI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4-8, 2020, Proceedings 1*. Springer International Publishing, 2020.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-60334-2_32

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Automatic Detection of Neonatal Brain Injury on MRI." *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis: First International Workshop, ASMUS 2020, and 5th International Workshop, PIPPI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4-8, 2020, Proceedings 1*. Springer International Publishing, 2020.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-60334-2_32

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Automatic Detection of Neonatal Brain Injury on MRI." *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis: First International Workshop, ASMUS 2020, and 5th International Workshop, PIPPI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4-8, 2020, Proceedings 1*. Springer International Publishing, 2020.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-60334-2_32

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Automatic Detection of Neonatal Brain Injury on MRI." *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis: First International Workshop, ASMUS 2020, and 5th International Workshop, PIPPI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4-8, 2020, Proceedings 1*. Springer International Publishing, 2020.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-60334-2_32

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Automatic Detection of Neonatal Brain Injury on MRI." *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis: First International Workshop, ASMUS 2020, and 5th International Workshop, PIPPI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4-8, 2020, Proceedings 1*. Springer International Publishing, 2020.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-60334-2_32

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Automatic Detection of Neonatal Brain Injury on MRI." *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis: First International Workshop, ASMUS 2020, and 5th International Workshop, PIPPI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4-8, 2020, Proceedings 1*. Springer International Publishing, 2020.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-60334-2_32

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Automatic Detection of Neonatal Brain Injury on MRI." *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis: First International Workshop, ASMUS 2020, and 5th International Workshop, PIPPI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4-8, 2020, Proceedings 1*. Springer International Publishing, 2020.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-60334-2_32

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Automatic Detection of Neonatal Brain Injury on MRI." *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis: First International Workshop, ASMUS 2020, and 5th International Workshop, PIPPI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4-8, 2020, Proceedings 1*. Springer International Publishing, 2020.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-60334-2_32

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Automatic Detection of Neonatal Brain Injury on MRI." *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis: First International Workshop, ASMUS 2020, and 5th International Workshop, PIPPI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4-8, 2020, Proceedings 1*. Springer International Publishing, 2020.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-60334-2_32

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Automatic Detection of Neonatal Brain Injury on MRI." *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis: First International Workshop, ASMUS 2020, and 5th International Workshop, PIPPI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4-8, 2020, Proceedings 1*. Springer International Publishing, 2020.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-60334-2_32

8.2 Chapter 4 Publication - Detection of Injury and Automated Triage of Preterm Neonatal MRI Using Patch-Based Gaussian Processes

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Detection of Injury and Automated Triage of Preterm Neonatal MRI Using Patch-Based Gaussian Processes." *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*. Springer International Publishing, 2021.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-87735-4_22

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Detection of Injury and Automated Triage of Preterm Neonatal MRI Using Patch-Based Gaussian Processes." *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*. Springer International Publishing, 2021.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-87735-4_22

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Detection of Injury and Automated Triage of Preterm Neonatal MRI Using Patch-Based Gaussian Processes." *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*. Springer International Publishing, 2021.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-87735-4_22

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Detection of Injury and Automated Triage of Preterm Neonatal MRI Using Patch-Based Gaussian Processes." *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*. Springer International Publishing, 2021.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-87735-4_22

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Detection of Injury and Automated Triage of Preterm Neonatal MRI Using Patch-Based Gaussian Processes." *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*. Springer International Publishing, 2021.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-87735-4_22

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Detection of Injury and Automated Triage of Preterm Neonatal MRI Using Patch-Based Gaussian Processes." *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*. Springer International Publishing, 2021.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-87735-4_22

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Detection of Injury and Automated Triage of Preterm Neonatal MRI Using Patch-Based Gaussian Processes." *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*. Springer International Publishing, 2021.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-87735-4_22

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Detection of Injury and Automated Triage of Preterm Neonatal MRI Using Patch-Based Gaussian Processes." *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*. Springer International Publishing, 2021.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-87735-4_22

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Detection of Injury and Automated Triage of Preterm Neonatal MRI Using Patch-Based Gaussian Processes." *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*. Springer International Publishing, 2021.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-87735-4_22

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Detection of Injury and Automated Triage of Preterm Neonatal MRI Using Patch-Based Gaussian Processes." *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*. Springer International Publishing, 2021.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-87735-4_22

This section has been redacted for copywrite reasons.

Citation:

Macleod, Russell, et al. "Detection of Injury and Automated Triage of Preterm Neonatal MRI Using Patch-Based Gaussian Processes." *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*. Springer International Publishing, 2021.

Publication link:

https://link.springer.com/chapter/10.1007/978-3-030-87735-4_22

8.3 Chapter 5 Raw Data Tables

Table 8.1 Sensitivity and specificity scores for each HC vs HIE classification models using raw Z-score histogram from each modality.

Model	T1w	T2w	Combined	ADC
LR	0.774/0.827	0.839/0.808	0.919/0.846	0.855/0.577
LIN-SVM	0.806/0.865	0.839/0.750	0.871/0.827	0.839/0.962
RBF-SVM	0.758/0.115	0.726/0.115	0.774/0.115	0.855/0.865
RF	0.871/0.808	0.806/0.635	0.855/0.750	0.839/0.885
GPC	0.855/0.865	0.855/0.673	0.871/0.846	0.919/0.904

Table 8.2 Permutation testing of the HC vs HIE classification models using raw Z-score histograms from each modality. Each cell contains the model accuracy and random permuted accuracy (chance) on the first line and FDR corrected p-scores on the second.

Model	T1w	T2w	Combined	ADC
LR	0.540/0.836 0.003	0.540/0.793 0.002	0.526/0.860 0.004	0.533/0.765 0.001
LIN-SVM	0.529/0.837 0.003	0.532/0.756 0.001	0.524/0.825 0.002	0.539/0.876 0.007
RBF-SVM	0.554/0.630 0.091	0.541/0.581 0.157	0.531/0.521 0.153	0.536/0.841 0.003
RF	0.536/0.814 0.002	0.532/0.727 0.001	0.538/0.803 0.002	0.529/0.867 0.005
GPC	0.531/0.850 0.004	0.532/0.798 0.002	0.532/0.861 0.004	0.530/0.896 0.013

Table 8.3 Sensitivity and specificity scores for each HC vs HIE classification models using optimised regional abnormal voxel count from each modality.

Model	T1w	T2w	Combined	ADC
LR	0.790/0.827	0.871/0.827	0.676/0.712	0.855/0.885
LIN-SVM	0.806/0.808	0.806/0.827	0.794/0.750	0.935/0.923
RBF-SVM	0.726/0.885	0.790/0.615	0.647/0.731	0.839/0.865
RF	0.919/0.846	0.887/0.788	0.912/0.865	0.887/0.846
GPC	0.855/0.846	0.823/0.827	0.765/0.808	0.919/0.885

Table 8.4 Permutation testing of the HC vs HIE classification models using optimised regional abnormal voxel count from each modality as input. Each cell contains the model accuracy and random permuted accuracy (chance) on the first line and FDR corrected p-scores on the second.

Model	T1w	T2w	Combined	ADC
LR	0.532/0.829 0.002	0.539/0.830 0.002	0.531/0.745 0.001	0.540/0.884 0.010
LIN-SVM	0.521/0.810 0.002	0.533/0.756 0.001	0.529/0.778 0.001	0.532/0.923 0.040
RBF-SVM	0.533/0.781 <i>0.001</i>	0.540/0.769 0.001	0.524/0.707 0.001	0.529/0.821 0.002
RF	0.539/0.871 0.006	0.538/0.832 0.002	0.540/0.879 0.008	0.528/0.850 <i>0.004</i>
GPC	0.529/0.842 0.003	0.532/0.828 0.002	0.533/0.781 <i>0.001</i>	0.536/0.907 0.020

Table 8.5 Sensitivity and specificity scores for each HIE normal vs poor outcome classification models using raw Z-score histogram from each modality.

Model	T1w	T2w	Combined	ADC
LR	0.412/0.600	0.588/0.600	0.471/0.600	0.471/0.533
LIN-SVM	0.647/0.600	0.412/0.533	0.471/0.600	0.529/0.467
RBF-SVM	1.000/0.000	1.000/0.000	1.000/0.000	1.000/0.000
RF	0.529/0.533	0.588/0.600	0.588/0.600	0.706/0.733
GPC	0.421/0.533	0.737/0.533	0.474/0.667	0.529/0.667

Table 8.6 Permutation testing of the HIE normal vs poor outcome classification models using raw Z-score histograms from each modality. Each cell contains the model accuracy and random permuted accuracy (chance) on the first line and FDR corrected p-scores on the second.

Model	T1w	T2w	Combined	ADC
LR	0.486/0.514	0.476/0.662	0.483/0.596	0.485/0.501
	0.488	0.171	0.303	0.498
LIN-SVM	0.479/0.586	0.479/0.532	0.478/0.556	0.476/0.503
	0.307	0.455	0.410	0.504
RBF-SVM	0.481/0.478	0.477/0.478	0.479/0.478	0.477/0.611
	0.531	0.532	0.517	0.355
RF	0.477/0.520	0.480/0.579	0.479/0.588	0.483/0.591
	0.477	0.322	0.314	0.317
GPC	0.480/0.491	0.478/0.612	0.474/0.592	0.481/0.603
	0.509	0.295	0.330	0.317

Table 8.7 Sensitivity and specificity scores for each HIE normal vs poor outcome classification models using raw Z-score histogram from each modality.

Model	T1w	T2w	Combined	ADC
LR	0.588/0.933	0.471/0.867	0.529/0.533	0.471/0.867
LIN-SVM	0.765/0.800	0.412/0.933	0.647/0.400	0.706/0.600
RBF-SVM	0.588/0.533	0.529/0.867	0.647/0.067	0.588/0.400
RF	0.471/0.933	0.824/0.733	0.588/0.267	0.765/0.733
GPC	0.353/0.933	0.647/0.733	0.471/0.667	0.412/0.600

Table 8.8 Permutation testing of HIE normal vs poor outcome classification models using optimised regional abnormal voxel count from each modality. Each cell contains the model accuracy and random permuted accuracy (chance) on the first line and FDR corrected p-scores on the second.

Model	T1w	T2w	Combined	ADC
LR	0.483/0.715 0.092	0.479/0.632 0.247	0.479/0.541 0.430	0.477/0.625 0.257
LIN-SVM	0.484/0.734 0.090	0.481/0.639 0.237	0.483/0.529 0.452	0.479/0.672 0.156
RBF-SVM	0.479/0.548 0.431	0.479/0.707 0.097	0.479/0.478 0.504	0.483/0.486 0.515
RF	0.486/0.736 0.089	0.483/0.789 0.072	0.476/0.493 0.515	0.486/0.750 0.085
GPC	0.477/0.685 0.133	0.481/0.691 0.132	0.483/0.511 0.485	0.480/0.545 0.428