KTH ROYAL INSTITUTE
OF TECHNOLOGY

Degree Project in Medical Engineering

Second cycle 30 credits

Estimating Brain Maturation in Very Preterm Neonates

An Explainable Machine Learning Approach

PATRIK SVENSSON

Stockholm, Sweden 2023

KTH ROYAL INSTITUTE OF TECHNOLOGY

Estimating Brain Maturation in Very Preterm Neonates

An Explainable Machine Learning Approach

PATRIK SVENSSON

Master's Programme, Medical Engineering, 120 credits
Date: June 15, 2023

Supervisor: Raúl Benítez
Reviewer: Xiaogai Li
Examiner: Matilda Larsson
School of Engineering Sciences in Chemistry, Biotechnology and Health
Host organization: Universitat Politècnica de Catalunya
Swedish title: Estimering av hjärnmognad i mycket prematura spädbarn
Swedish subtitle: En ansats att tillämpa förklarbar maskininlärning

Abstract

Introduction: *Assessing brain maturation in preterm neonates is essential for the health of the neonates. Machine learning methods have been introduced as a prospective assessment tool for neonatal electroencephalogram (EEG) signals. Explainable methods are essential in the medical field, and more research regarding explainability is needed in the field of using machine learning for neonatal EEG analysis.*
Methodology: *This thesis develops an explainable machine learning model that estimates postmenstrual age in very preterm neonates from EEG signals and investigates the importance of the features used in the model. Dual-channel EEG signals had been collected from 14 healthy preterm neonates of postmenstrual age spanning 25 to 32 weeks. The signals were converted to amplitude-integrated EEG (aEEG) and a list of features was extracted from the signals. A regression tree model was developed and the feature importance of the model was assessed using permutation importance and Shapley additive explanations.*
Results: *The model had an RMSE of 1.73 weeks (R^2=0.45, PCC=0.676). The best feature was the mean amplitude of the lower envelope of the signal, followed by signal time spent over 100 μV.*
Conclusion: *The model is performing comparably to human experts, and as it can be improved in multiple ways, this result indicates a promising outlook for explainable machine learning model applications in neonatal EEG analysis.*

Keywords

Preterm Neonates, Brain Maturation, aEEG, Explainable Machine Learning, Feature Importance

Abstract

Introduktion: *Att bedöma hjärnmognaden hos för tidigt födda spädbarn är väsentligt för barnets hälsa. Maskininlärningsmetoder har introducerats som verktyg för att analysera neonatala elektroencefalogram(EEG)signaler. Förklarbara metoder är essentiella i det medicinska fältet, och mer forskning behövs kring förklarbara metoder i neonatal EEG-analys.*
Metodologi: *Detta examensarbete tar fram en förklarbar maskininlärningsmodell som estimerar postmenstrual ålder hos väldigt prematura spädbarn från EEG signaler och undersöker vilka signalegenskaper som är av högst vikt för modellens estimering. Tvåkanaliga EEG-signaler hade samlats in från 14 friska spädbarn med postmenstrual ålder mellan 25 och 32 veckor. Signalerna konverterades till amplitudintegrerad EEG (aEEG) och en lista med signalegenskaper extraherades. En regressionsträdsmodell utvecklades och signalegenskapernas vikt bedömdes med permutationsvikt och Shapley additive explanations.*
Resultat: *Modellen hade ett RMSE på 1.73 veckor (R^2=0.45, PCC=0.676). Signalegenskapen som var viktigast för modellen var medelvärdet av den undre enveloppen följt av hur mycket tid signalen var över 100 μV.*
Slutsats: *Modellen estimerar åldern ungefär lika bra som experter, och då modellen kan förbättras på flera sätt indikerar detta resultat en lovande framtidsutsikt för förklarbara modellers användning inom neonatal EEG-analys.*

Nyckelord

Prematura spädbarn, Hjärnmognad, aEEG, Förklarbar maskininlärning

Acknowledgments

I would like to thank my supervisor Raúl Benítez for his positive energy and inspiration and for giving insightful perspectives and guidance on the project.

I would also like to thank Dr Ana Alarcon Allen and Mar Velilla Aparicio for helping me gain an understanding of how the measurements are taken in the hospital, what the neonatologists look at in the signal and giving helpful clinical insight to me coming from a more technical background.

Lastly I would also like to thank my friends and family for their support throughout the project.

Barcelona, June 2023
Patrik Svensson

Contents

List of Figures

List of Tables

List of acronyms and abbreviations

aEEG Amplitude-Integrated EEG

AI Artificial Intelligence

EEG Electroencephalography

HFD Highuchi Fractal Dimension

ML Machine Learning

NICU Neonatal Intensive Care Unit

PCC Pearson's Correlation Coefficient

PMA Postmenstrual Age

PSD Power Spectral Density

RMSE Root Mean Square Error

SD Standard Deviation

SHAP Shapley Additive Explanations

XAI Explainable Artificial Intelligence

Chapter 1

Introduction

More than 13 million babies were born preterm worldwide in 2020 [1], and preterm birth complications are the leading cause of death for children under 5 years of age [2]. A preterm newborn's health is even more fragile than a term newborn's and it is essential for the well-being of the infant to detect abnormalities as early as possible. Monitoring certain postnatal signals is helpful to determine whether the infant needs additional intervention [3]. Different technical aids have been developed for neonatologists to help monitor and take care of infants to prevent injury and death.

Monitoring electroencephalography (EEG) signals after birth is a common method for getting insights into the health of preterm neonates in neonatal intensive care units (NICUs). EEG signals often have artefacts and noise, and reading EEG signals properly and making an accurate estimation is a skill that takes time to develop. Clinicians with the required amount of experience to relatively accurately interpret EEG signals might not always be available when needed, and machine learning methods are investigated as an aid in the interpretation with promising results [4–6]. Neonatologists often monitor a processed version of the EEG called amplitude-integrated EEG (aEEG), and this is used in some NICUs for clinical assessment of the maturation. The visual assessment of the aEEG or EEG is a subjective measure and the result can vary noticeably even among expert neonatologists [7].

Using machine learning models could avoid the effect of the subjective nature of human visual EEG assessment, and could potentially produce consistent and accurate results. These models usually take in so-called features calculated from the data instead of taking the unprocessed data as input. The quality and relevance of extracted features are highly correlated with the result of the model, but it is not always clear what features are

of importance in a signal as signals can be complex. Investigating which features have the highest correlation with accurate predictions is of interest for improving the performance of these models, which in turn will help make robust and effective machine learning methods for aiding neonatologists in making accurate assessments.

In the medical field machine learning models must be explainable, meaning that it is possible for the clinicians using them to understand how it works and see how the model came to its conclusion. If non-explainable models are used in clinical practice, then the clinicians would not know how a model came to its conclusion, and basing a medical decision on an unexplainable measurement is a risk for the clinician who is responsible as well as for the patient. More information is needed in the area of explainable medical machine learning regarding neonates and EEG signals.

This work will focus on assessing the maturation of the brain of the infant from the aEEG and the end aim of the study is to answer the questions:

1. **Using an explainable machine learning model, which aEEG signal features (of the ones investigated) are of the highest relevance for estimating postmenstrual age in preterm infants?**
2. **How well can a simple explainable machine learning model predict postmenstrual age in preterm infants from aEEG data?**

A limitation is that it is not viable to investigate all possible features as it is possible to construct any number of features. How the features were chosen is detailed in section 3.

If a model can correctly estimate the brain age in healthy subjects from the aEEG, then it would be possible to apply it to signals from a new subject to see if and how much the subject is deviating in maturation. In real clinical practice, clinicians know the brain's age and try to estimate whether the brain is as mature as it should be. In this project, only signals from healthy subjects are used so that the model is trained as a reference. A subject is considered healthy if the subject grew up without neurological pathological conditions and maturational deviations.

The long-term vision is to contribute to a field of knowledge that will help produce effective and explainable machine learning models that could be used widely in NICUs to help preterm neonates grow up healthy.

Structure of the Thesis

The thesis will first introduce the project, then present the necessary background information as well as related works in the field, followed by a method

chapter detailing how the practical part was carried out. Then the results will be presented and discussed, and the limitations and ethical aspects will also be discussed. Lastly, the conclusions are presented together with ideas for future work.

Chapter 2

Background

This chapter will first summarize the general theoretical and practical concepts behind the project and then bring up similar works within this field of research.

2.1 Electroencephalography

Electroencephalography (EEG) is a non-invasive method for measuring electric potential differences in the brain. The signals are recorded by placing multiple electrodes on the scalp of the patient and measuring the voltage difference between the electrodes. It has been shown that early deviations in the EEG signal in neonates are associated with delayed cognitive development [8]. Measuring EEG is of clinical importance as it can help find developmental deviations and it is a method that is widely used in neonatology intensive care units (NICUs) for continuous evaluation of newborn children [7].

An EEG signal normally consists of multiple voltage recordings from different combinations of electrodes placed on the scalp, see figure 2.1 for an example of a recording. There is a standardized configuration for the placement of the recording electrodes called the 10-20 system, see figure 2.2. There are different types of electrodes available and different methods of how to place them and which placements in the 10-20 system to use if not all placements are used. The number of electrodes can also vary, and fewer electrodes are generally used when the patient is a neonate, as the head size is smaller [11] which reduces the number of electrodes that is possible to place.

Examples of signal characteristics that can be investigated are envelope characteristics, thresholds, the characteristics of so-called bursts (in periods when they occur, the recording in figure 2.1 is during a period with bursts), spectral characteristics such as the signal energy in different frequency bands

Figure 2.1: Image over multi-channel EEG. Burst onsets are indicated by a black arrow and burst ends are indicated by a white arrow. Image distributed under creative commons license [9].

(it is common to investigate the delta, theta, alpha and beta intervals which are the intervals 0.5-4 Hz, 4-7 Hz, 7-13 Hz and 13-30 Hz respectively), entropy and spectral slope.

2.1.1 Amplitude-Integrated Electroencephalography

Neonatologists often look at the so-called amplitude-integrated EEG (aEEG) which is a version of the raw EEG which makes it possible for neonatologists to look at longer periods of data at the same time compared to traditional EEG, commonly hours or days. While it is not a substitute for the raw EEG signal, it does provide an easier overview and insight into the EEG for visual inspection. The aEEG is created from the EEG by applying a band-pass filter between 2 Hz and 15 Hz of high order [12]. There are low-frequency artefacts lower than 2 Hz and there are frequencies for muscle activity and other physiological artefacts higher than 15 Hz and both are unwanted in the signal. The filter is asymmetric and has a slope in the frequency domain, where the higher cutoff frequency has a higher amplitude to compensate for the lower frequency content in the signal in the higher frequencies. The slope in frequency is 12dB/decade [12]. The signal is rectified after the filtering which means that

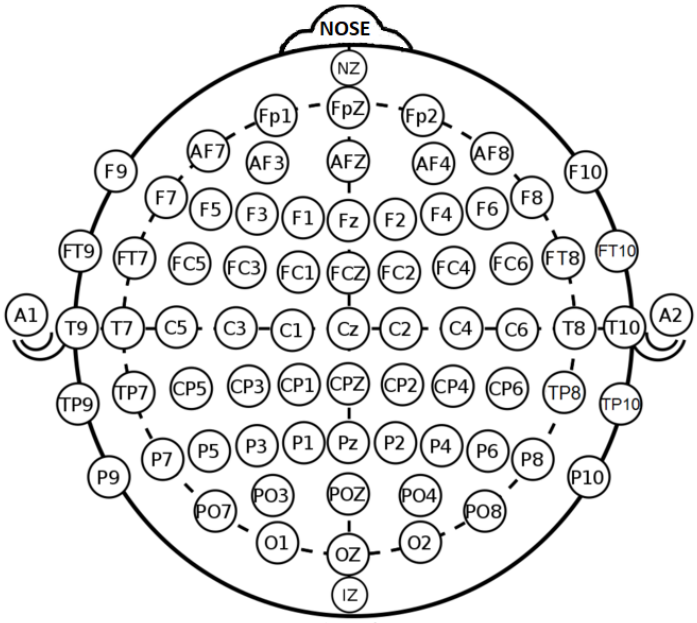

Figure 2.2: The international 10-20 system and the names of the positions. Top-down view of the head. Image distributed under creative commons license [10].

the absolute value y of the signal x

$$y[n] = |x[n]| \tag{2.1}$$

is taken at each sample n. This is a non-invertible operation which affects the frequency components in the signal in a non-linear way. This makes the frequency components of x inaccessible when only given y.

After the rectification, the envelope of the signal is produced by finding all local maxima of the signal and connecting them using interpolation. The envelope is then time-compressed before plotting. The aEEG is in other words a time-compressed envelope of the filtered and rectified EEG signal. In the resulting plot, the y-axis of the aEEG is linear between 0-10 μV and a logarithmic axis between 10-100μV.

The signal-to-noise ratio in EEG signals is often low compared to other physiological electric measurements like electrocardiography and electromyography, making the useful information in the collected signal harder to distinguish from the noise. Also, placing the electrodes correctly is a difficult task that requires substantial skill [11] which results in that the signal can contain artefacts or periods of disconnection that impair the interpretation of the signal. A noisy signal with periods of disconnection can present problems both when interpreting the signal manually and when using digital methods. While an experienced clinician can mostly distinguish a noisy signal from a clean signal, care has to be taken when developing digital methods as they need to have robust handling of noise, artefacts and disconnects.

However, technological advances in the field of machine learning give promising opportunities to develop digital methods to interpret the signal. When it comes to drawing conclusions about specific questions from the EEG, machine learning methods have been shown to perform equally or better than experienced clinicians [5, 6, 13].

2.2 Machine Learning

Machine learning (ML) is a field of finding and using methods that improve when they are exposed to data relevant to the task they are to perform. The design of an ML model allows the model to change some of its parameters when it is exposed to the data, which means that it improves itself. This can for instance be used to give insight into the data or perform tasks more effectively than humans. The methods are usually applied to specific tasks. ML is a type of artificial intelligence (AI).

ML models usually take so-called features as inputs. A feature can take different forms but in this project it is a scalar that is extracted from the data, for instance mean value or standard deviation. Any scalar that can be constructed from the data can be a feature, and there are infinite possibilities when choosing which features to extract from the data. The model in this project takes these features as input; it does not take the raw data as input.

2.2.1 Explainable Machine Learning

The field of explainable AI (XAI) is a developing branch of AI research which aims to develop methods and approaches that will make the models understandable by humans [14]. XAI is seeing increased interest following the implementation of the General Data Protection Regulation (GDPR) in the European Union, as users have now the right to know how a decision was made from an ML system [15, 16]. XAI is of especially high interest in the medical field as transparency is needed to make the methods trustable [17].

Developing effective ML methods in healthcare is of high relevance for streamlining procedures in healthcare and lowering costs, and interest in AI methods in the medical field has risen in recent years [14]. But in the field of medical ML, the models need to be explainable to make the models trustable by the clinicians [16]. It is up to the clinician to make the final medical decision, and blindly trusting the conclusion of a model is a risk for the patient [18], and can pose a juridical risk to the responsible clinician as well. Therefore it needs to be possible to inspect how the model came to its conclusion, and developing medical ML models that are explainable is key for them to be used and implemented in the medical field [16]. If it is possible to see how the model came to its conclusion, then the decision flow can be seen and the model's conclusion validated or discarded by the clinician. Explainability is also important when using AI in a clinical setting to not interfere with ethical principles and explainability is a requirement in medical AI [19].

Explainability is an aspect of AI models which assesses how well the model itself can be explained and if its outcomes can be understood, reproduced and traced through the model. Explainability encompasses the model, the input data and the interaction with the human user. It takes into consideration the goal of the human user. If a model has high explainability, then the user has a good understanding of the inner workings of the model and how the model reaches the output result from the input data [20].

XAI models can be distinguished into two types; post-hoc and ante-hoc

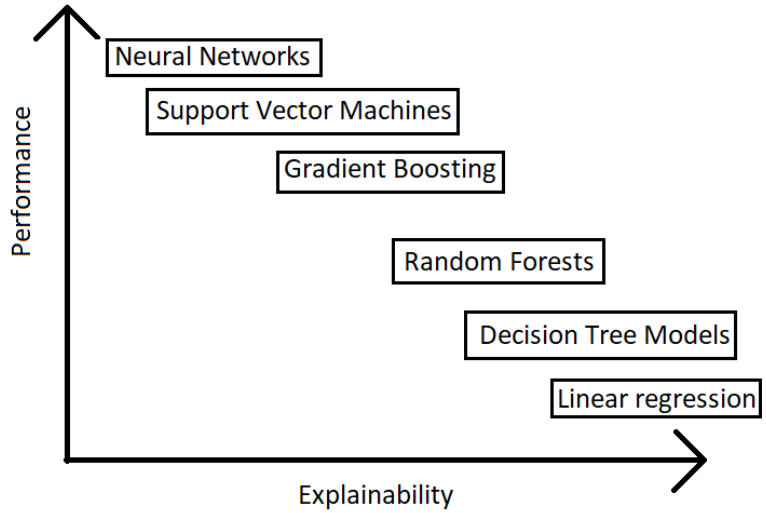

Figure 2.3: The tradeoff between performance and explainability.

explainable. The post-hoc explainable systems are systems which can provide explanations for a specific decision and the decision can be reproduced, but it is not possible to provide explanations of the whole system [16]. So-called black box models like neural networks generally fall into this category.

Ante-hoc systems are transparent and explainable by design and can be called glass box systems. Examples of ante-hoc systems are linear regression and fuzzy interference systems [16]. Machine learning models that are simple are usually more transparent, for instance models based on linear regression [14].

The unexplainable models usually perform better than the explainable ones and there is a tradeoff between performance and explainability, see figure 2.3. This introduces some tension in the field of AI, regarding if explainability is more important than performance. In the field of medical AI however, explainability is essential [19, 21]. Using black-box models can violate patients' right to autonomy and informed consent [22].

2.2.2 Regression Trees

One type of ante-hoc, glass box explainable ML model is regression trees [16]. They are based on decision trees which are widely used data structures for making decisions and predictions, see figure 2.4 for a simple example of a decision tree.

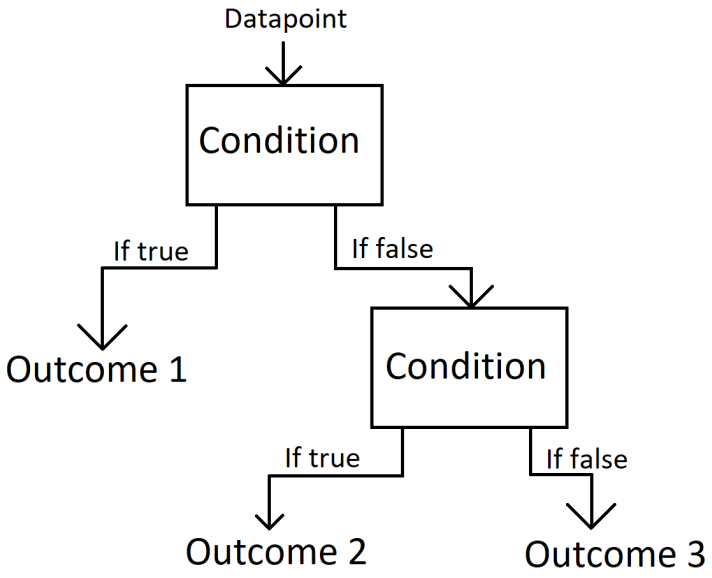

Figure 2.4: A simple decision tree.

Starting from the top and going down in the decision tree in figure 2.4, the data available gets split in each condition node depending on the condition. This condition is based on a calculated feature from the data (for example: "Is the value of feature X in this datapoint higher than 10?"), and it is checked if the condition is true or false. The features of each datapoint in the dataset have been calculated before applying the decision tree model to the data.

When developing a decision tree with ML methods, a dataset which includes the correct value of the target variable in each datapoint is needed to train the model. The dataset is split into two parts, one called the training set and the other called the test set. The training set is usually larger and has the correct value of the target variable kept in each datapoint. In the test part of the dataset, the value of the target variable is omitted. The training data is then given to the model. The model uses this training data set to create the condition nodes in the tree. When this is done then the nodes in the tree do not change, and the model is said to be fitted, or trained. The end nodes in the tree are called leaf nodes.

Then the test dataset, where the datapoints' correct labels are omitted from the model, is given to the model and each datapoint is assessed and predictions are made for the target variable. A comparison can then be made between the predicted outcome of the fitted model and the correct one, and the model's accuracy can thus be assessed. During training the nodes in the tree change,

but the nodes do not change after the training is finished. So when the model makes predictions, it is easy to see the path that each datapoint has taken, and the decisions made by the decision tree are fully transparent.

2.2.3 Evaluating Model Performance

The performance of ML models based on regression is commonly evaluated through the root mean square error (RMSE) and the R^2-value. The root mean square error is defined as

$$RMSE = \sqrt{\frac{\sum_{k=1}^{K}(Predicted_k - Label_k)^2}{K}} \tag{2.2}$$

where K is the total number of datapoints in the test set, $Predicted$ is the predicted outcome from the model and $Label$ is the actual value of the target variable. RMSE gives the error in the same unit as the target variable and is calculated on the model's performance on the test set.

The R^2 value is a relative measure of how well the model fits the data. It is unitless with a value up to 1 where 1 is a perfect fit to the data. It is defined as

$$R^2 = 1 - \frac{\sum_{k=1}^{K}(Predicted_k - Label_k)^2}{\sum_{k=1}^{K}(Label_k - \bar{L})^2} \tag{2.3}$$

where $\bar{L}$ is the mean of the labels. If the R^2 is less than 1 it means that not all of the variability of the data can be explained by the model and, for instance, a R^2 of 0.75 means that 25% of the variability in the data cannot be explained by the model [23]. A model that always predicts $\bar{L}$ will have a R^2 of 0 and if it performs worse than that then the R^2 will be negative.

2.2.4 Overfitting

When training a model it is important to not train the model to be too fitted to the training dataset; it is normally desired to have a model with some generalizability so it also will perform well on data points that it has not been trained on. When a model is too specific to the training dataset it is said to be overfitted. Preventing overfitting is considered when designing the model, i.e. choosing its parameters, based on the training set at hand as well as the nature

of the data and how the model would behave when applied to novel data.

2.3 Features and Feature Importance

For ML models, the choice of features directly impacts the performance of the model. To optimize the performance of the model, it is important to accurately extract relevant features from the data. More information is needed in this field about which features in the aEEG will lead to good performance when using explainable ML models.

Assessing which features are of the highest importance for the outcome is of relevance for developing effective ML models, especially for decision tree-based models as they are prone to overfitting if a large number of features are used [24].

Feature importance also provides insight into the data. It could indicate that one type of feature is more relevant than another type of feature, which gives information about what is important for estimating the target variable. It could then be investigated further to find the optimal feature in that feature group for improving the model's performance.

Two different methods are used in this project for assessing feature importance: Permutation feature importance and Shapley additive explanations (SHAP) feature importance and the underlying theory is briefly explained here.

2.3.1 Permutation Feature Importance

Permutation feature importance is one method to quantify the features' importance on a model's output. The method is a measurement based on the decrease in the model's performance when shuffling the features one at a time [24].

The method works by taking the fitted model as input, as well as data which can be either training data or test data. It then computes a reference score s_{ref} using the R^2 value by inputting the data to the model without alternating any of the features. Then it shuffles one feature, meaning that it takes all the values of one feature in the dataset and randomly rearranges them, therefore breaking its relationship with the target variable. This results in a corrupted version of the original dataset which is then inputted into the fitted model and the R^2-score is calculated. This is repeated K times and the importance of the feature i_f is then calculated as the reference score minus the mean of the scores of the

corrupted datasets [24]:

$$i = s_{\text{ref}} - \frac{1}{K}\sum_{k=1}^{K} s_k \tag{2.4}$$

This is then repeated for each feature in the dataset to extract the importance value for each feature.

If $K = 1$ then the results can vary greatly due to the nature of randomness when shuffling. Higher K therefore stabilize the measurement but also increases computational cost [25].

Permutation feature importance is model agnostic, meaning that it can be applied to any ML model, but it is also model specific, meaning the results will only be relevant for that specific model [24]. The method only gives information about which features affect the model output and how much, but not in which direction each feature affects the output.

2.3.2 Shapley Additive Explanations

Another method to assess feature importance is by using Shapley additive explanations (SHAP), presented in [26]. The method builds on Shapley values, presented in [27], which are based on game theory. An intuitive explanation of the Shapley values is that they treat all the features as players in a game. The output of the model is the result of the game, and the feature importance is regarding how much each feature contributed to the result of the game. The method determines this by introducing the features randomly into the coalition of features already in the game and assesses the average change in result for the coalition when the new feature is introduced. The average change in output is the Shapley value for the feature.

The SHAP method then uses these values to build an additive explanation model g as

$$g(c) = s_0 + \sum_{m=1}^{M} s_m c_m \tag{2.5}$$

where $c \in \{0, 1\}^M$ is the coalition vector (describing which features are in the game), M is the total number of features, s_m is the Shapley value for feature m and s_0 is the chance output of the original model, meaning the predicted output if none of the features has any impact on the output. This is an additive feature attribution method which means it attributes an effect to each feature, which approximates the output of the original model when all feature attributes are

summed.

The magnitude and sign of each feature attribute give information both about how much each feature affects the model's output, as well as if it affects the output to be higher or lower; SHAP feature importance results are directional [26]. The Shapley values are also unique and there are no sources of randomness affecting them so the results will always be the same.

SHAP is generally computationally heavy but the authors of the SHAP paper also presented a version of the method specifically for models based on decision trees called TreeSHAP, presented in [28]. The version makes the method less computationally heavy when applied to tree-based models, as the time complexity of the method is decreased from exponential to polynomial. For a complete mathematical formulation of the SHAP method, see the original paper [26], and for the TreeSHAP see [28].

2.3.3 The Importance of Accurate Importance Methods

In the article [28] the authors apply different commonly used feature importance methods (Saabas, gain, split, permutation and SHAP) on simple regression trees and point out that permutation-based importance and SHAP importance are the only consistent ones; the other methods are not. Inconsistency here means that if a model is changed so that it relies more on a given feature, the feature importance methods could estimate the feature as being less important than before the change. This hinders the ability to compare different methods, and also weakens the trustworthiness of the feature importance results, as a feature with a higher connection to the output prediction can be considered less important than another feature with a lower connection to the output prediction. As the field of explainable models becomes increasingly important, and as feature-importance methods seek to explain models accurately, it is essential to use consistent feature-importance methods.

Both permutation feature importance and SHAP are model agnostic and can be applied to any type of ML model. The achieved results after applying the methods to a model are however tied to the model it has been applied to.

2.3.4 Correlated Features

If some features are correlated then the feature importance result can be misrepresentative because if a signal characteristic has high relevance for

model prediction output but that characteristic is represented by multiple features which are heavily correlated, then all of those features will have a lower feature importance result than if only one of them were used. This can lead to a misrepresentation of the feature importance result, and therefore it is of relevance to find correlated features and either exclude some of them or merge them together in a relevant way [25]. Selecting features based on correlation is used in biomedical data analysis [29].

2.3.5 Brief Explanation of Each Feature

A brief explanation of the features investigated in this project will be presented here.

TIME DOMAIN FEATURES

Envelope Features

The envelope features are constructed by extracting the lower and upper envelopes of the aEEG signal and then taking the mean of the samples of the envelope. The mean value of the lower envelope gives an approximation of how strong the weaker parts of the signal are. This is something that the clinicians visually inspect [30]. The upper envelope was also included, as well as the difference between them. The difference indicates how wide the signal is in the time domain.

The variability of the lower part of the signal is also visually inspected in clinics so the standard deviation of the low envelope was also extracted.

Time Above Thresholds

The thresholding features consist of the time the signal spends above a certain voltage threshold. There are four of them and they register how much time the signal is above 10 µV, 25 µV 50 µV and 100 µV respectively.

Higuchi Fractal Dimension

The Higuchi fractal dimension (HFD) is a non-linear method originating from chaos theory for measuring signal complexity [31]. There are other ways to calculate fractal dimensions but Higuchi's method is the most accurate [32].

The HFD is calculated by first defining lengths

$$L_m(k) = \frac{N-1}{(\frac{N-m}{k})k^2} \sum_{i=1}^{\frac{N-m}{k}} |s_N(m+ik) - s_N(m+(i-1)k)| \tag{2.6}$$

where N is the number of samples, s is the signal, m is the initial time and k is the interval time between samples. The term before the summation is a normalization factor. Then the sum of the lengths between each of the samples in the signal is taken, followed by defining the length

$$L(k) = \frac{1}{k} \sum_{m=1}^{k} L_m(k) \tag{2.7}$$

from which the following points are calculated

$$\Big(\log \frac{1}{k}, \log L(k) \Big) \tag{2.8}$$

and the HFD is then the slope of the best-fitting linear function passing through the points [33, 34].

HFD is a relevant measurement to include in automated signal analyses as it is an important measurement within medical research. Its use has risen within medical research, including in neurophysiology and EEG studies [31]. HFD has an advantage over linear measurements as physiological signals are often nonstationary and nonlinear, and HFD is a good numerical measurement no matter if the signal is stationary, nonstationary, stochastic or deterministic [31].

FREQUENCY DOMAIN FEATURES

Mean PSD - Signal Energy

The power spectrum density (PSD) of a signal is a measure based on the Fourier transform of the signal. It presents the signal's power in different frequencies instead of over time. The sum of the values in the PSD gives the total energy in the signal, and then dividing by the number of samples gives the mean signal energy value per sample.

Spectral Entropy

Spectral entropy is a way to quantify the irregularity of a signal by using information theory and measuring the amount of information needed to describe the distribution of the PSD.

First the PSD needs to be calculated from the signal and then spectral entropy S_e is calculated as

$$S_e = -\sum_{k=0}^{f_s/2} P[k] \ \log_2(P[k]) \tag{2.9}$$

where f_s is the sampling frequency, $P[k]$ is the value of the PSD at discrete frequency k [35].

A high entropy value corresponds to an unpredictable signal, as more information is needed to describe it, while a low entropy corresponds to a more predictable and regular spectrum.

Spectral Flatness (Wiener entropy)

The spectral flatness S_f is defined as the geometric mean of the PSD of a signal and divided by its arithmetic mean

$$S_f = \frac{g}{a} \tag{2.10}$$

where g is the geometric mean and a is the arithmetic mean, i.e.

$$g = \left(\prod_{k=0}^{M} P[k] \right)^{\frac{1}{M}} \tag{2.11}$$

and

$$a = \frac{1}{M} \sum_{k=0}^{M} P[k] \tag{2.12}$$

where $P[k]$ is the value of the PSD at discrete frequency k and M is the total number of samples in the PSD [36]. The spectral flatness indicates how flat the PSD is. It approaches a value of 1 for white noise and approaches 0 for a single pure sine wave.

2.4 Related Works

Dong et al. [37] developed a machine-learning model based on a gradient-boosted method and applied it to a dataset of EEG recordings collected from 1851 neonates in a NICU to estimate brain age. They used a total of 59 features extracted from the signals. The results showed that their model can achieve a Pearson's correlation coefficient(PCC) of 0.904 between its prediction and the real outcomes and concluded that their model can successfully predict brain maturation and that the model could benefit clinicians working in the NICU. Gradient boosted methods are however generally not explainable, see figure 2.3.

Stevenson et al. [7] measured the accuracy and reliability between experts (with 16 years of mean experience in reading EEG charts) assessing the EEG data and experts assessing the aEEG data of preterm neonates. They found that the experts assessing the aEEG had a higher rate of agreement than the group of experts assessing EEG, while the group of experts assessing the EEG had a lower systematic error (EEG overestimated 0.8 weeks compared to aEEG underestimating 1.8 weeks from actual postmenstrual age). The random error was 1.7 weeks and 1.8 weeks for the EEG and aEEG, respectively).

They also compared the visual performance of the experts to a computational measure of the postmenstrual age, using calculations based on the EEG and a machine learning model. The computational model's result had both a lower systematic error (underestimating by 0.1 weeks) and a lower random error (1.1 weeks) compared to the two expert groups. They conclude that automated computational measures may be more accurate than human experts for the estimation of brain maturity. The computations were based on a support vector regression model, which is generally seen as not explainable.

The authors also tested their ML model on data from another dataset that the model has never seen before, and then the model performed comparably to the experts.

O'Toole et.al. [5] developed a model for brain maturation estimation in the very and extremely preterm neonate using a 9-electrode-EEG. They used 41 features extracted from the data. The authors investigated inputting different sets of features into their model and noted that using the model on multiple features yielded better results than only using a single feature. The best feature combination yielded a result with a mean square error of 82 days, which corresponds to an RMSE of 1.294 weeks. This is a lower error than their reported chance reading which has an RMSE of 2.33 weeks.

Stevenson et.al. [6] used a 9-electrode-EEG to estimate maturational age in

preterm infants using a support vector regression model using 23 features. The data was recorded from 39 infants and consisted of 567 one-hour long epochs after removing epochs due to artefacts. The correlation between their model's prediction and the clinically determined maturational age was $PCC = 0.936$. They also investigated which features showed a significant correlation with age, and for the full EEG the significant features were the 95% percentile of the reference EEG, 95% percentile of the envelope, total spectral power and the number of bursts per hour.

While these related works have shown that it is possible to use ML methods to achieve better results than clinicians, none of them has used explainable ML models. Unexplainable ML models are not likely to be adopted for widespread use in NICU as most clinicians prefer, and medical ethics demand, that the methods be explainable so it is possible to see on what grounds a decision was made [19]. And this is where this project comes in.

Chapter 3

Methodology

The general idea of the method was to take the recorded dataset and process the signals, extract features, develop a model, and then use the feature importance methods with the model to find the best features which is the main result of the study. Lastly a model is produced to assess the performance of an explainable ML model based on the features.

The pipeline was programmed in Python and the package for the machine learning models was scikit-learn [24]. Other packages used were shap [26], seaborn [38], NumPy [39] and pandas [40].

The full signal flow of the data is presented at the end of the chapter in figure 3.9.

3.1 Signal Processing

3.1.1 The Dataset

The data in the dataset was collected before the start of this project from 14 neonates in the NICU at the Sant Joan de Déu Hospital in Barcelona. The group that collected the data has ethics approval, and the data was collected with consent from the parents of the study subjects. The data was anonymized and shared, with approval, with the lab where this project was conducted. The data is part of an ongoing prospective study from the research institute Institut de Recerca Sant Joan de Déu.

The data was collected using the OBM Olympic Brainz monitor (Natus Medical Inc., San Carlos, CA, USA) and adhesive electrodes placed on the scalp of the neonates. The electrodes were placed in 10-20 system positions C3, P3, C4 and P4 and the collected signals were in the pairs C3-P3 and C4-

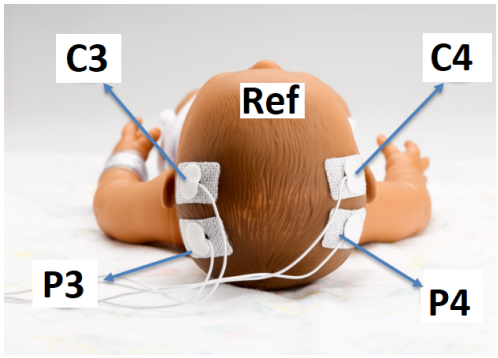

Figure 3.1: The placements of the electrodes.

P4, see figure 3.1. A reference electrode was also placed on the forehead. The impedance for each electrode was also recorded. Each measurement was 185 minutes long.

All the subjects were healthy, were not on medication and grew up without neurological pathological conditions and without deviations from the maturational process in the brain. This is essential when creating a model to later compare novel measurements to. The training data must be representative of a healthy population. If the training dataset had patients with patients deviating from the normal progress or possibly having a disease that affects the EEG then the model would be trained incorrectly. The age distribution of the neonates can be seen in figure 3.2. The age is the postmenstrual age PMA, which is the age from the first day of the mother's last menstrual period. The dataset consists of very preterm neonates as they are aged under 32 weeks, except for one neonate with an age of 32.1 weeks [41].

3.1.2 Signal Pre-Processing and Segmentation

The raw signals were first subjected to artefact removal using the electrode impedance signals (if the impedance in an electrode was higher than a certain value, then that part of the signal was replaced with the global mean of the signal) and an artefact removal method based on Gaussian mixture models that checks for outliers by creating Gaussian distributions and removing samples that fall outside of the distributions. The EEG signals were then converted into aEEG by filtering, rectification and extraction of the envelope as described in section 2.1.1. The time-compression part of the aEEG is just a matter of how to display the signal; no signal processing part was needed to perform the time-compression. This was however done before and not part of the work in this

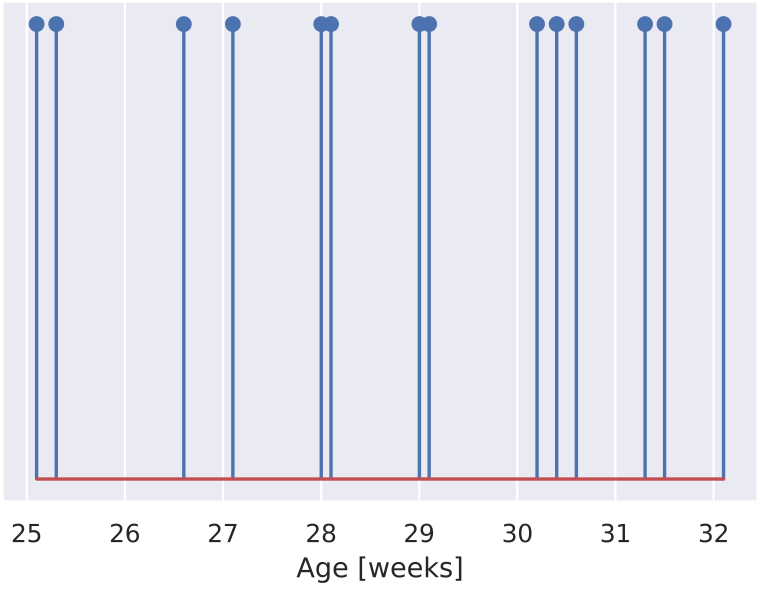

Figure 3.2: Stem plot over the distribution of the PMA of the 14 neonates in the dataset.

project.

The dataset was considered too small with only 14 signals so some measures needed to be taken to portion the dataset into more datapoints.

While it would be possible to use each whole signal of 185 minutes to extract the features, it was decided to portion the signals into 10-minute segments. This divides it into more datapoints and give more weight to a statistical assessment of the model's performance. If the model performs well it also means that only a 10-minute long signal is necessary to estimate the brain maturation of the neonate. The disadvantage of doing this is that signal characteristics that appear over a longer period would not be included and the relatively short length of the segment might however affect the performance of the model negatively.

The segmentation was made with a sliding window with 50% overlap, see figure 3.3, which gave more segments than just slicing the signal into 10 min long pieces.

It was also decided to uncouple the right and left EEG-channel to double the number of segments, and no differentiation was made after this point between if the signal originated from the left or right half of the brain. This was deemed acceptable because of the erratic nature and similarity in signal behaviour between the two channels. This was also deemed acceptable

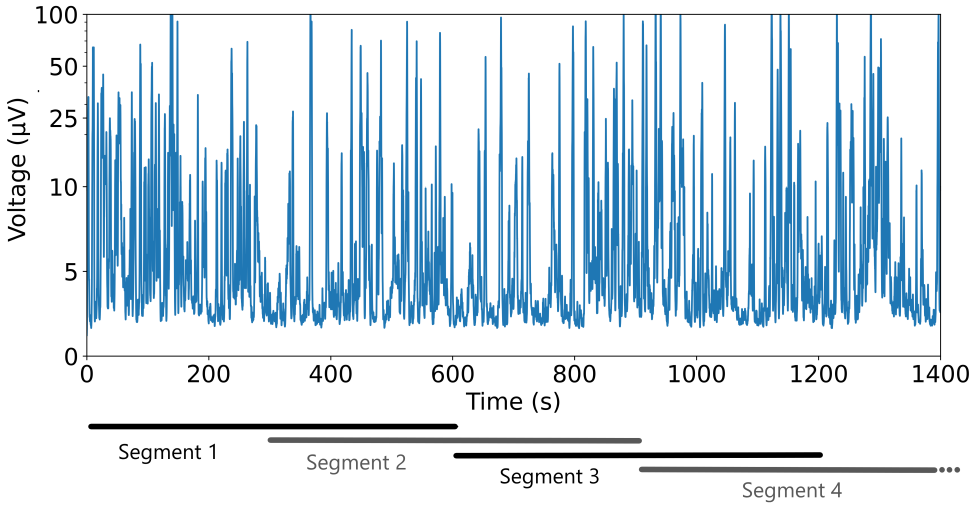

Figure 3.3: Figure showing how the segmentation was performed.

because the correlation between the same type of features in the right and the left signal was high, which doesn't give much more new information and is also undesirable in feature importance methods. Splitting them gave more segments with original values as well as avoiding the correlation problem. This choice also changed the design in such a way that the algorithm can work on data from only one channel which means that the algorithm can be applied even in a case when one of the electrodes has disconnected, or when only one channel EEG is used, which potentially could ease the electrode placement in the NICU. It does however mean that two-channel features like cross-correlation or brain symmetry index cannot be used in the model.

After applying these methods, the dataset consisted of a total of 862 segments.

3.1.3 Removal of Faulty Segments

Due to the inherent prevalence of artefacts and disconnections in EEG signals, each segment was checked to make sure they contained reasonable values. If the mean of the 5% of the samples with the highest value were higher than 125 µV, then the segment was considered faulty and removed due to strong artefacts. It was also removed if the mean value of the 5% of the lowest samples were under 1.5 µV due to disconnects. The threshold values were found by iteratively changing the threshold values and visually inspecting the segments

that get removed. It was tuned from loosely set thresholds to more and more strict thresholds until a few segments that visually appeared normal started to get removed. It was deemed better to have strict thresholds, i.e. rather remove a few normal segments than missing to remove a few with artefacts or disconnects. This is because it is hard to predict how faulty data affects the model and feature importance analysis; it was deemed preferable to have a clean dataset to improve the trustworthiness of the result. Examples of signals that were picked up by the filter and removed can be seen in figure 3.4.

3.1.4 Feature Extraction

A list of signal features was formulated, based on signal features being investigated in research.

- Mean value of lower envelope
- Mean value of upper envelope
- Difference between mean of upper and lower envelope
- Standard deviation (SD) of lower envelope
- Time above 10 µV
- Time above 25 µV
- Time above 50 µV
- Time above 100 µV
- Higuchi fractal dimension
- Spectral entropy
- Spectral flatness
- Mean PSD amplitude

The features were selected with the following criteria:

- Single channel feature (i.e. not cross-correlation or similar).
- Possible to represent the characteristic of interest in a 10 min segment.

The features are mostly time domain features, as the rectification of the signal when transforming it to aEEG breaks the relationship between the frequencies in the original signal and the aEEG. So investigating some frequency-based features like the prevalence of signal components in the delta, theta, alpha and beta frequency intervals is not useful.

Two example signals can be seen in figure 3.5, one from a more premature neonate (25.3 weeks old) and one from a less premature neonate (32.1 weeks old). These specific segments were chosen as examples as they were quite

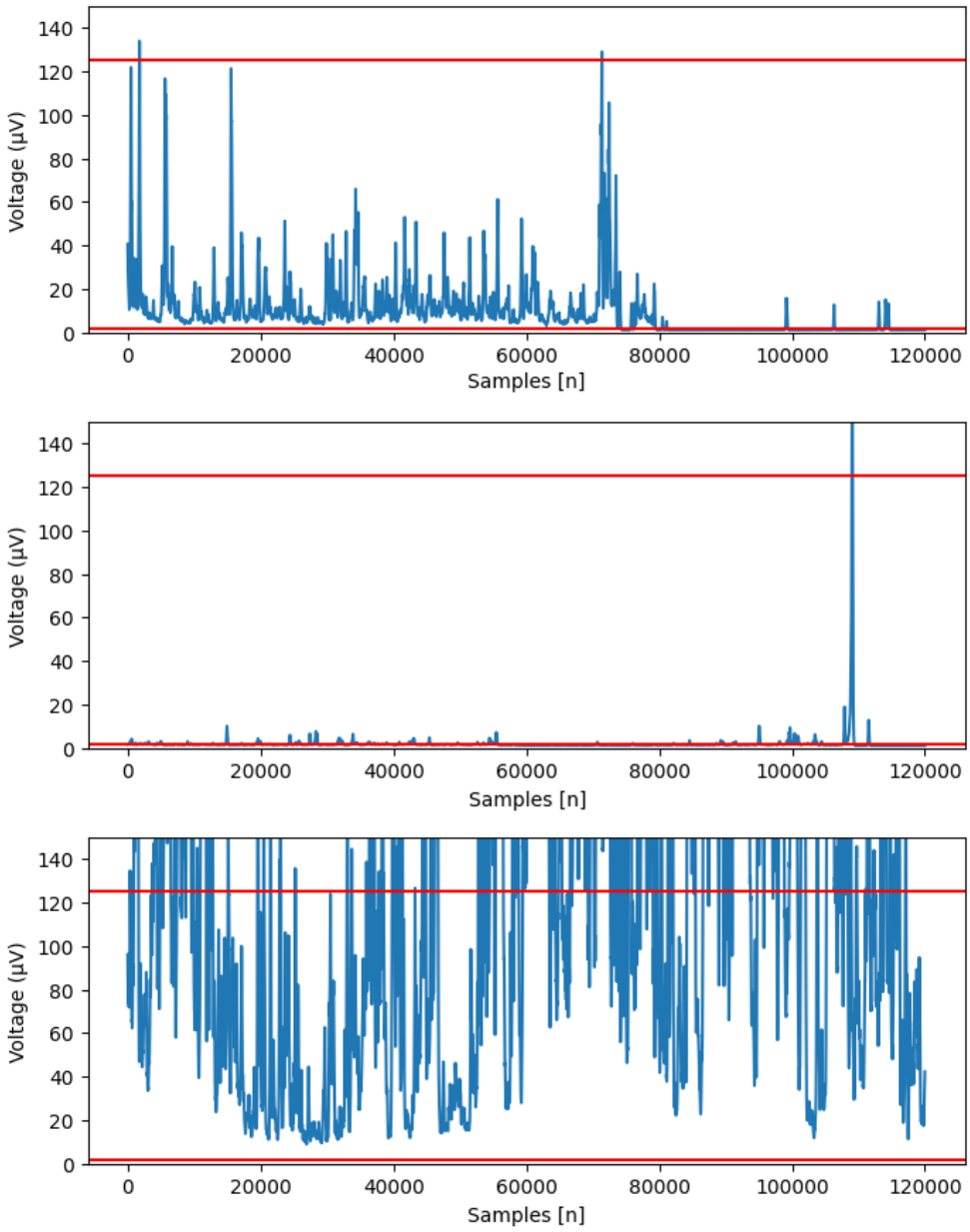

Figure 3.4: Three example segments that were removed based on the threshold filter. The thresholds can be seen as red lines. Approximately the first 8000 samples of the top signal look normal but then an electrode likely disconnected from the scalp. The second is a segment during disconnect; the dataset contained multiple of these. In the third signal there are plenty of artefacts and the signal does not behave as expected.

representative of the signals from the ages of the respective neonates. In the younger neonate it can be seen that there seems to be a vague lower baseline of the signal around 2-3 μV, but it can also be seen that the voltage often spikes up higher. For the older neonate, a lower baseline cannot be seen, and the whole signal seems to fluctuate approximately around 10 μV. These are some of the factors that the choice of features was based on. The vague lower baseline could be represented by the mean value of the lower envelope, and the difference between the mean of the upper and lower envelope could represent the breadth of the signal which seems to differ between different levels of maturity in the neonates. The standard deviation of the lower envelope was also included as there seems to be a higher standard deviation in the older neonate, and it could be a good indicator of maturation.

The features based on the time the signal is above certain thresholds were chosen because it seemed to be a good way to capture the signal differences between less mature and more mature neonates. While the signal from the more mature neonate in figure 3.5 seems to spend more time above 10 μV than the younger neonate, the less mature neonate's signal spikes up above 50 μV more often. A few thresholds were therefore decided and included as features.

Spectral entropy, spectral flatness, the mean PSD amplitude (mean sample signal energy) and Higuchi fractal dimension were also included in this study as these are features that are investigated in the literature and are used in unexplainable ML models that can predict maturation [5, 6, 31, 37].

The feature values for the two example signals in figure 3.5 can be seen in table 3.1. One thing that can be noted is that the signal of the younger neonate spends less time above 10 μV and 25 μV but more time above 50 μV and 100 μV than the signal of the older neonate. Results from this study and other studies.

After extracting the features, some segments were removed based on the values of the extracted features. If a feature in a segment had an unreasonably high or low value, the segment was removed. This was implemented as an extra measure to find faulty segments and the decision to include this was based on the artefact-prone nature of EEG signals. The thresholds for inclusion was the mean value of respective feature ± 3 times its standard deviation. This limit was decided upon after visually inspecting the datapoints and inspecting whether extreme outliers (which were faulty when inspecting the corresponding segment visually) were satisfactorily removed without removing datapoints with reasonable values. These thresholds were applied to most features. A few features had manual thresholds assigned when there

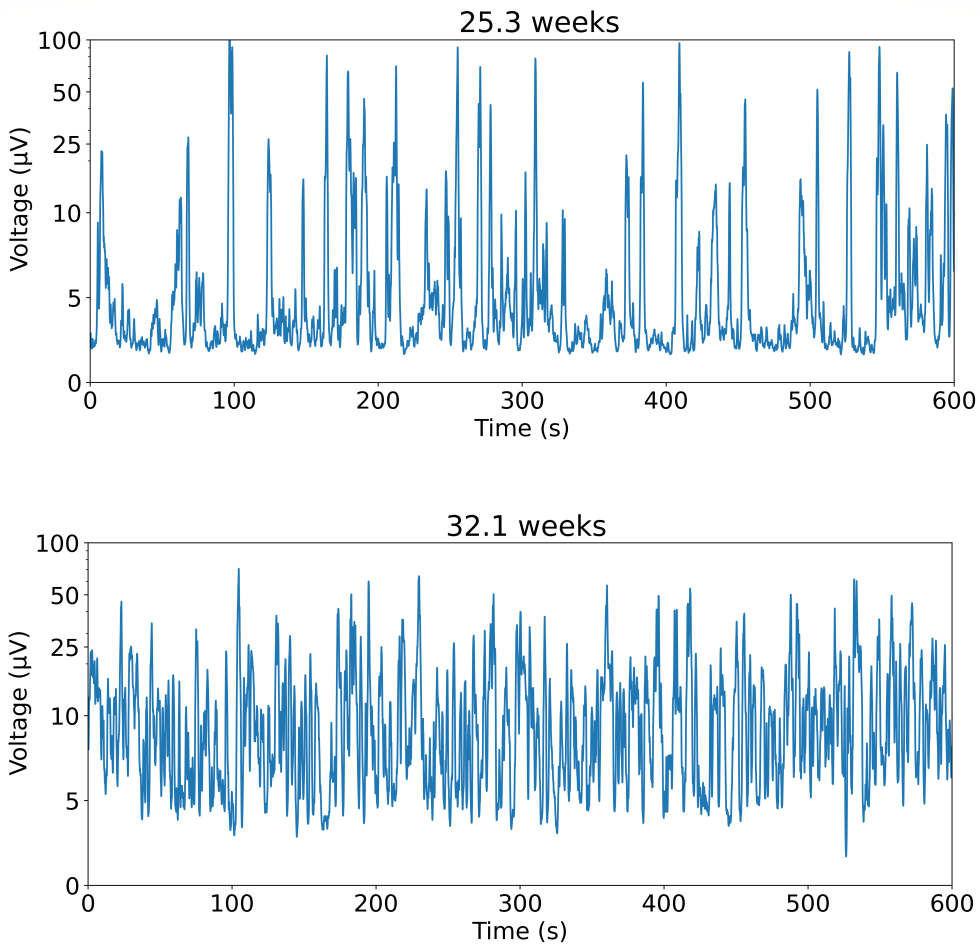

Figure 3.5: 10-minute segments from two different neonates. Note that the y-axis is linear up to 10 µV and logarithmic above.

Table 3.1: Feature values for the two example signals in figure 3.5.

Feature	**25.3 weeks**	**32.1 weeks**
Mean of lower envelope [µV]	4.40	10.0
Mean of upper envelope [µV]	6.02	12.5
Envelope mean difference [µV]	1.63	2.51
SD of lower envelope [µV]	7.06	7.81
Time above 10 µV [s]	75.2	258
Time above 25 µV [s]	25.3	42.3
Time above 50 µV [s]	9.64	3.16
Time above 100 µV [s]	0.4	0
Higuchi fractal dimension	1.0034	1.0047
Spectral entropy	1.20	1.25
Spectral flatness	1.20×10^{-9}	1.42×10^{-9}
Mean PSD amplitude [pV^2/Hz]	1.15	0.62

was a clear limit for what is reasonable, for instance should the value of HFD be between 1 and 2. After this operation, the dataset consisted of 830 datapoints.

3.1.5 Removing Correlated Features

Removing correlated features is important for the permutation importance methods to give correct results, so the features' correlation was inspected at this stage.

In figure 3.6 it can be seen that some features are heavily correlated, especially the envelope features and the threshold features. This was not unexpected.

The removal of correlated features followed the following algorithm:

1. Find the highest correlation value in the correlation matrix and its two corresponding features.
2. If the correlation is higher than 0.5: Sum up the absolute values of the two rows of correlation values, and remove the one with the highest sum to be removed.
3. Calculate new correlation matrix.
4. Repeat until there are no unmarked features with a correlation higher than 0.5.

Doing feature selection based on correlation is a justified choice in biomedical data analysis, as mentioned in subsection 2.3.4, but choosing a threshold

value to base the removal on is a simple method. Other more advanced methods like principal component analysis or XAICFS-BDA presented in [29] were looked into and considered but then deemed too time-consuming to implement considering the relatively short list of features. So a simple, iterative feature selection method based on pairwise correlation was designed and used. The value of 0.5 was considered an upper limit as a correlation of 0.5 can be considered moderate. The highest correlation between features after using 0.5 as the threshold and applying the method was 0.33 which was estimated to be low enough to not interfere with the result. Lowering the threshold below 0.33 would result in less than 5 features left, which was deemed too restrictive. In this case, the correlation threshold could have been chosen between 0.33 and 0.5 and the results would have been the same.

The features that were kept after this operation were:

- Mean value of the lower envelope
- Time above 100 μV
- Spectral entropy
- Spectral flatness
- Mean PSD amplitude

The correlation heatmap between the features that were kept can be seen in figure 3.7.

3.2 Regression Tree Model

The machine learning model used for the feature importance was a regression tree model. This type of model was chosen based on its explainability. It is known that it would likely perform worse due to the performance-explainability tradeoff as mentioned in subsection 2.2.1, but the research question necessitated that an explainable model be used. The regression tree model type was chosen from other explainable models as tree-based models are amongst the most explainable. This was done to increase this project's relevance for contributing to the development of a model that is practically applied and used in NICUs.

The tree was designed so that the leaves in the tree are not smaller than 20 datapoints, and the maximum depth of the tree was limited to 4, meaning that the path from the first node to a leaf is no longer than 4 conditional nodes.

Restricting the minimum leaf size and the maximum depth was done to reduce the risk of overfitting. A too small leaf size leads to overfitting of the model, and a too large size leads to worse performance. The leaf size was

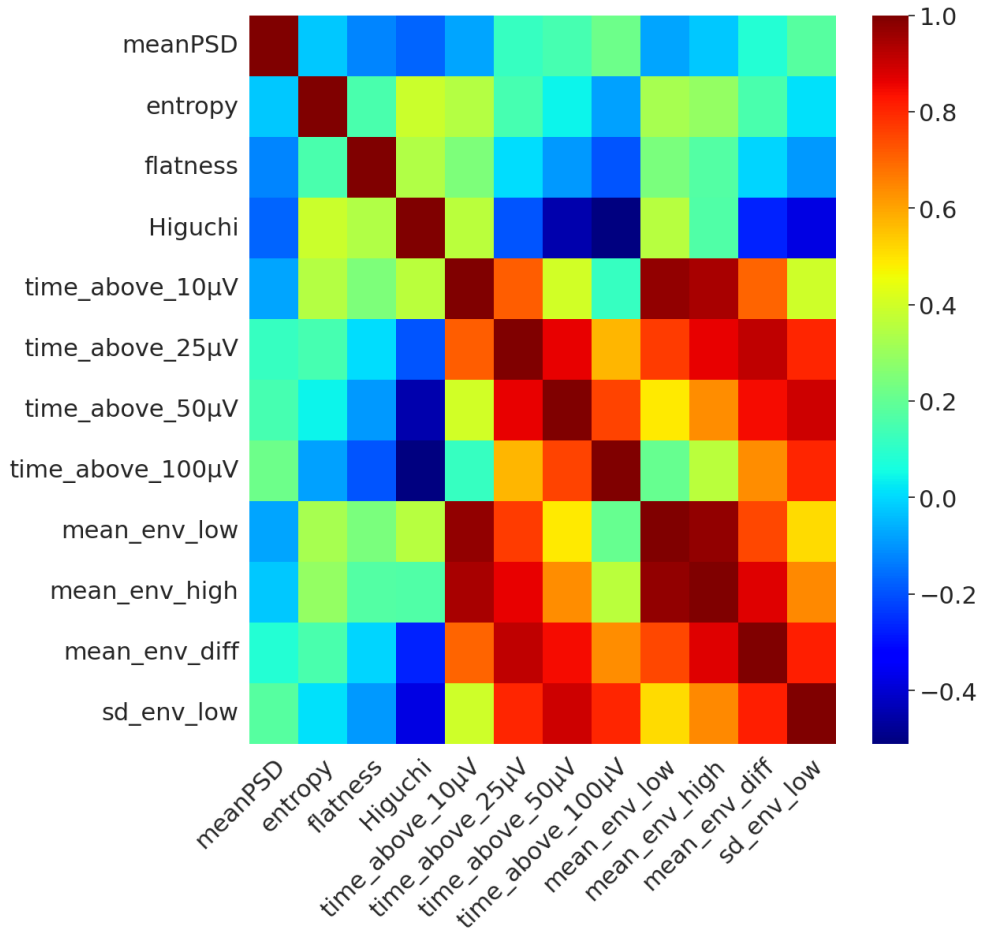

Figure 3.6: Correlation heatmap showing the correlation between all different features.

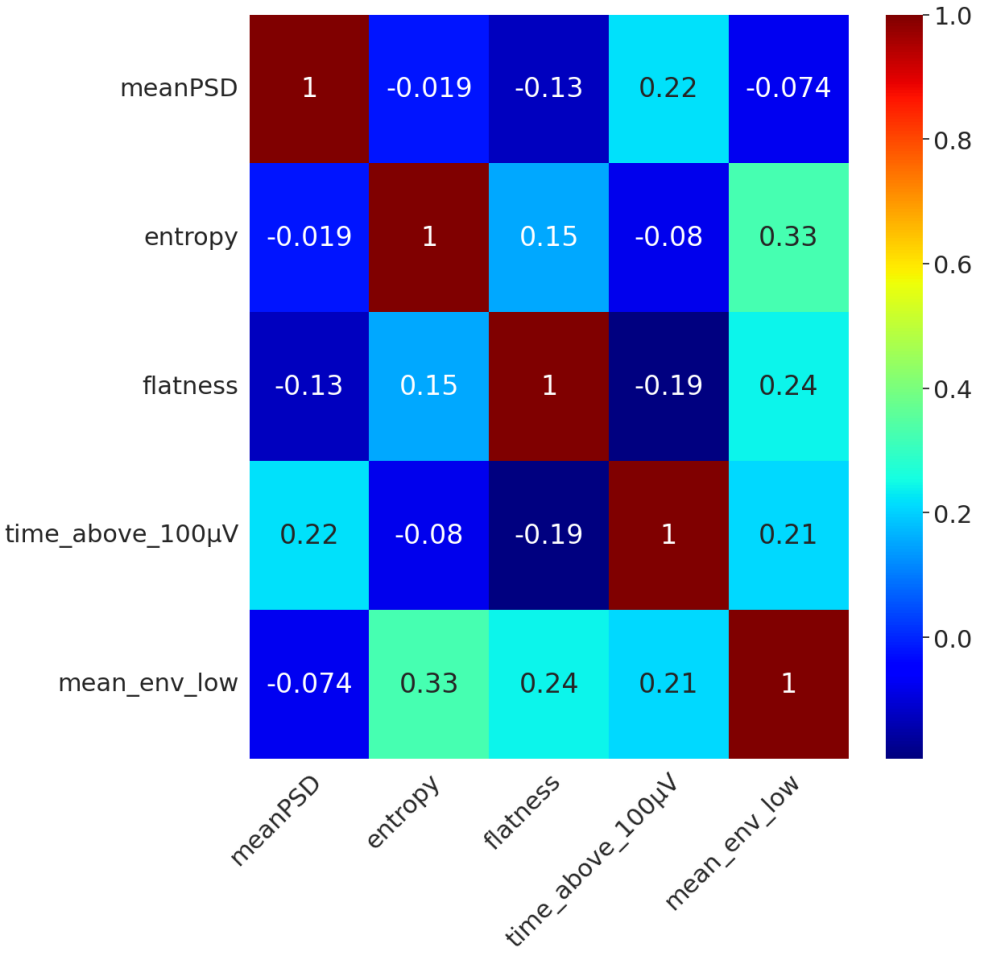

Figure 3.7: Correlation heatmap with correlation values showing the correlation between the features included in the model.

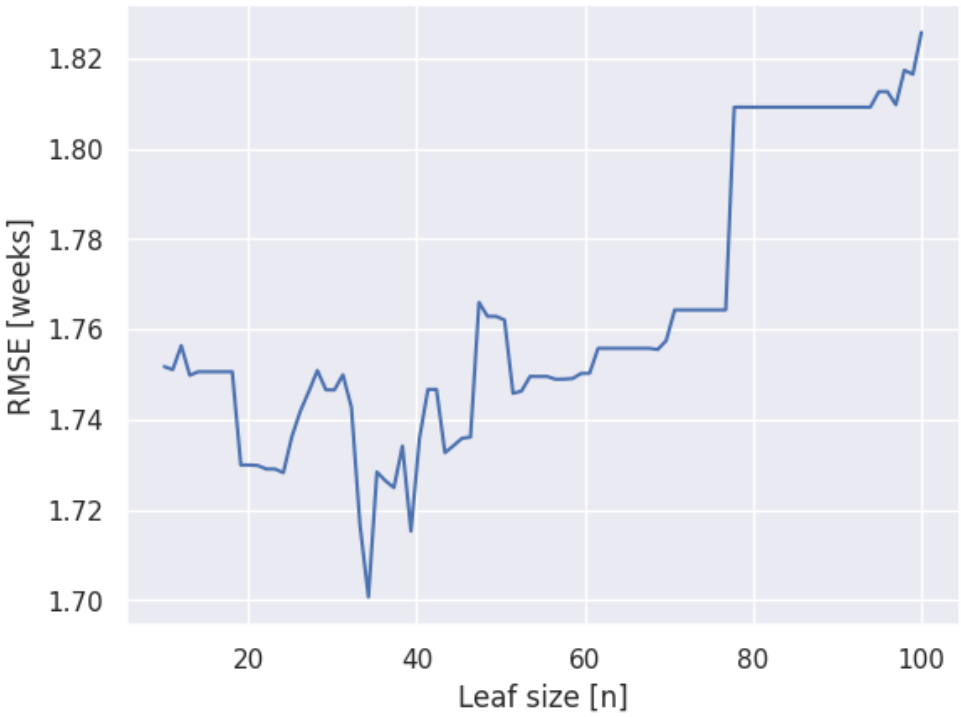

Figure 3.8: RMSE for different leaf sizes

investigated by calculating the RMSE of the model for different minimum leaf sizes and choosing an optimal value. A plot of how the RMSE varied over different leaf sizes can be seen in figure 3.8. The minimum leaf size was set to 20 as it minimizes the RMSE. This means that no leaf in the tree is smaller than 20. While it can be seen in the graph that a leaf size of 34 would generate an even lower error, the value 20 was chosen to be more representative, as this method has some random variability and the local minimum at 34 could be a result of that. The RMSE and R^2-value for this model were calculated for the model. PCC was also calculated between the predicted outputs and the real values.

The feature dataset was split into two cohorts. 80% of the data was for training the model, and the rest was for testing the model's accuracy. See figure A.1 in the appendix for the full regression tree.

3.2.1 Dummy Regressor Model

The model must perform significantly better than the chance level for feature importance methods. If the fitted model's performance is similar to the chance

level, then the feature importance method will not yield useful results [24]. Therefore it is important to evaluate the model's predictive power before performing feature importance methods.

After the model was created the model performance was assessed to be compared with the chance level as the performance must be higher than the chance level. Another regressor was created (DummyRegressor from scikit-learn) to calculate the chance level on the dataset, and it was then fitted to the same training data as the regression tree. The difference in performance between the dummy model and the regression tree model was regarded as big enough to perform the feature importance evaluation with results.

3.3 Feature Importance Evaluation

The feature importance was then made using two methods, permutation feature importance and SHAP feature importance.

The feature importance was performed on the test part and not the training part of the dataset. This gives information about how important the features are for the generalization of the model to novel data [24], which is of higher interest than how important the features are for the data the model is trained on. As is common with ML models, calculating results on data that the model has already seen is often too optimistic and not representative of the model's performance on novel data [25].

The permutation feature importance was then calculated using the scikit-learn package and the resulting graphs and values were extracted.

Then the feature importance using the SHAP method was calculated using the shap package. The importance was calculated on the test set for the same reasons as for the permutation importance and the results were extracted.

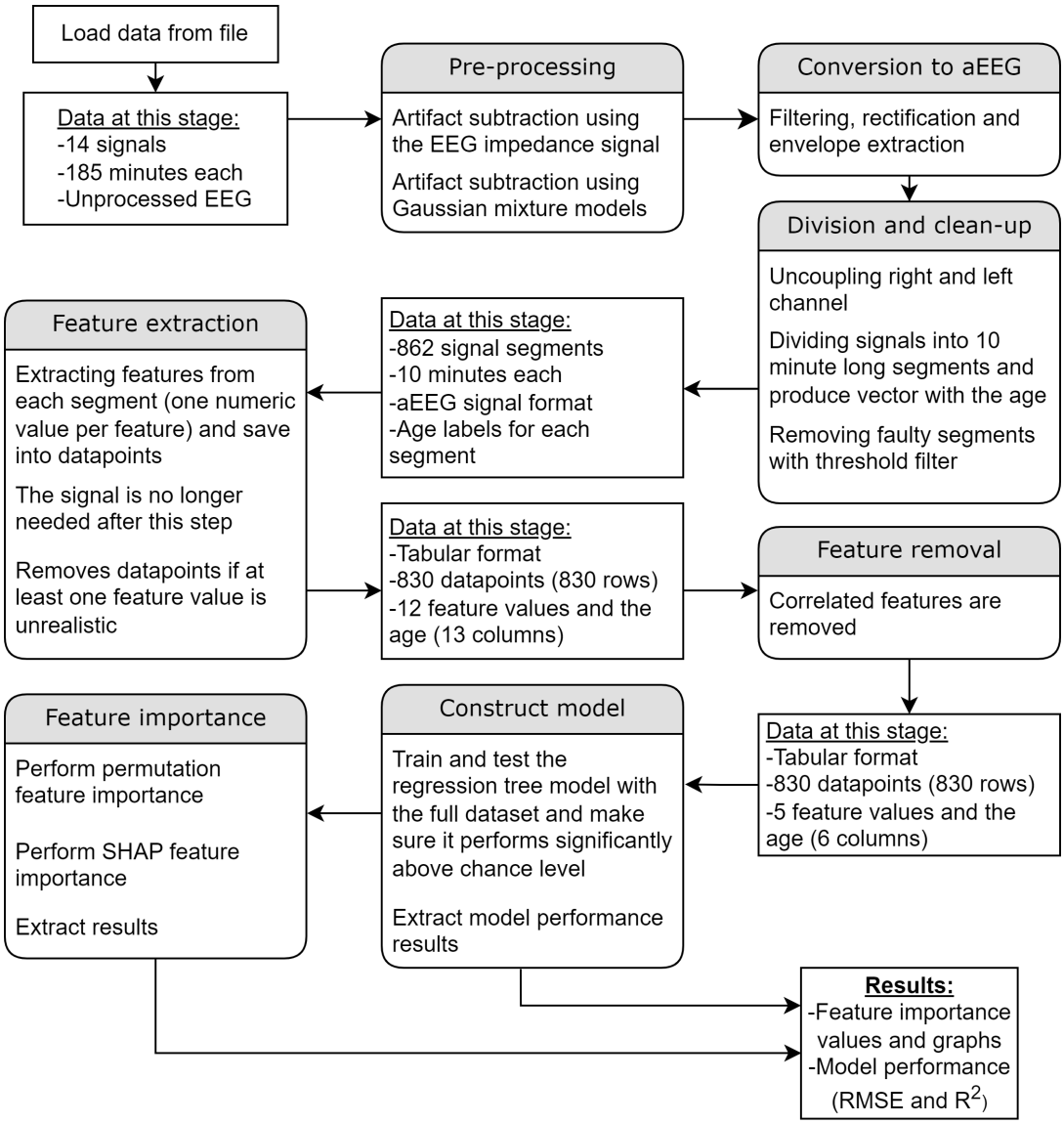

Figure 3.9: Full signal flow from raw data to the final result.

Chapter 4

Results

In this chapter the results from the project are briefly presented, to be further discussed in the next chapter.

4.1 Feature Importance

The permutation feature importance analysis showed that the mean of the

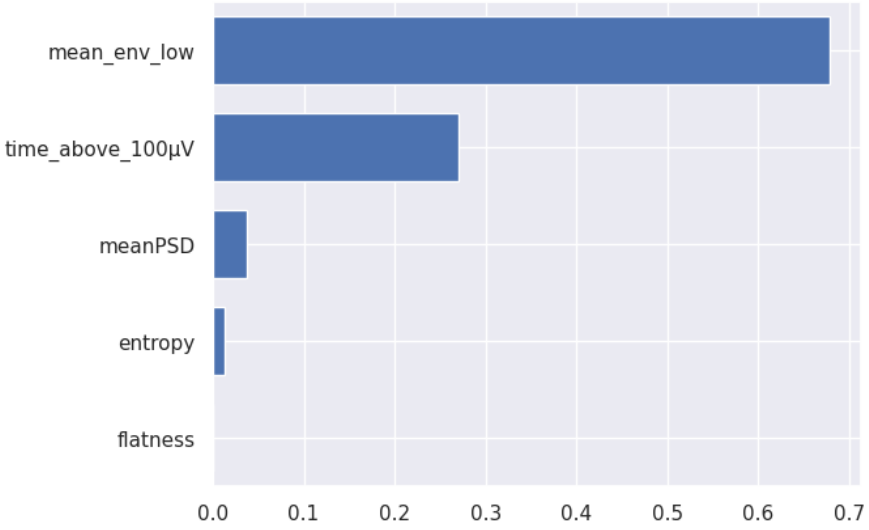

Figure 4.1: Permutation importance result

values of the low envelope is the most important feature for estimating brain maturation using the regression tree model. All of the features and their importance based on the permutation method can be seen in figure 4.1 and

Table 4.1: The permutation importance results in numbers.

Feature	Importance weight
Mean of low envelope	0.679
Time above 100 µV	0.270
Mean of PSD	0.038
Entropy	0.013
Spectral flatness	0

table 4.1. The staples in the graph sum up to 1, and the features make up for 0.68, 0.27, 0.04, 0.01 and 0 respectively. This means that the highest-ranked feature makes up for 68% of the predictive power in this model followed by 27%.

Figure 4.2: SHAP summary plot

The results from the importance analysis based on SHAP are shown in figure 4.2. The graph is ordered in regards to importance and it can be seen that this method ranked the features in the same order. The figure shows that a low value of the mean of the low envelope contributes to a lower output of the model, i.e. a lower age estimation. This is in accordance with the clinical assessment method as the stronger the lower part of the signal is, the more mature the brain is [42]. It can also be seen that a lower value of the time above 100 µV contributes to a higher maturation estimation. It can be seen that the result for the mean of the PSD affects the model's output less, but there seems to be a trend that a lower value of the PSD corresponds to a lower brain maturation; the signals' energy content is higher in more mature neonates.

Table 4.2: Correlation values between the two most important features and the removed features. Correlation above 0.5 is marked with an asterisk (*) and correlation above 0.9 is marked with two. The correlation between the two top features was 0.21.

	Low envelope mean	**Time above 100µV**
Time above 10 µV	0.97**	0.11
Time above 25 µV	0.77*	0.57*
Time above 50 µV	0.49	0.76*
High envelope mean	0.98**	0.36
Envelope mean difference	0.75*	0.63*
SD low envelope	0.51*	0.8*
HFD	0.35	-0.51*

This doesn't contribute much to the output prediction, but it does give some insight into the data. Both methods concluded that spectral flatness does not contribute to the output estimation of this model.

4.1.1 Correlated features

As the features that are correlated to the most important ones are also relevant for the prediction, they can be seen in table 4.2 together with the correlation value to the two most important features. When inspecting the aEEG, each feature presented by the feature importance results is relatively unrelated to other features and assesses different aspects of the signal, while the correlated features in the table will be different ways of looking at the same underlying variable.

Additionally, the table presents how all features have at least a 0.5 correlation to at least one of the two top features. This indicates that all the removed features do have some predictive power for the estimation.

The correlation between the mean of the low envelope and time above 10 µV and also the high envelope mean gives information that the latter two features would also contribute with similar predictive power to the model as the correlation is close to 1.

4.2 Model Performances

The performance of the model can be seen in table 4.3 together with the performance for the dummy model. An RMSE of 1.73 means that for any

Table 4.3: The performance of the model and the dummy model.

Model	**RMSE [weeks]**	$\boldsymbol{R^2}$	**PCC**
Regression tree	1.73	0.45	0.676
Dummy model	2.33	0.000018	-

given maturation estimation, it is likely to differ by 1.73 weeks from the real value. The R^2-value of 0.45 means that the model can explain about 45% of the variability in the data, meaning that it leaves 55% unexplained by the model.

As the two top features make up for 95% of the predictive power in the model and the model has an R^2 of 0.45, it means that the top two features account for 42% of correctly estimating brain age.

Compared to the result of the dummy model, the results show that the tree model performs 0.6 weeks better in general. The R^2 of the dummy model was expected to be close to 0, which it is. The dummy model generates a constant output; the estimation is the same for every datapoint, and therefore it is not possible to calculate PCC for the dummy model.

Chapter 5

Discussion

This chapter will discuss the results and their importance, as well as the limitations of the thesis and will briefly discuss the ethical aspects of this work.

5.1 Feature importance

The main purpose of this study was to investigate which features are of the highest importance for predicting brain maturation in pre-term neonates for an explainable ML model, and the results show that the mean value of the low envelope is the best predictor out of the features investigated, followed by the signal time over 100 µV. The low envelope mean made up 68% of the predictive power in the model and the time over 100 µV made up 27% of the predictive power. So the two most important features make up about 95% of the predictive power in the model. It can also be noted that both of the features are amplitude-based features.

These results indicate what features of the signal are correlated with brain maturation, and what features can be useful when developing explainable ML models in this field. However, as the feature importance method is model-specific, other results may be achieved if a different ML model is used.

In the SHAP results it can be seen that a high value of the mean of the lower envelope contributes to a higher output prediction, while a higher value of the time above 100µV contributes to a lower age. This indicates that the aEEG of a mature premature neonate has higher low parts of the signal, and doesn't spike up above 100 µV as the aEEG in a less mature neonate might do. From this it can be concluded that the trace of the lower part of the aEEG of a more mature neonate should be higher, and the signal should not often spike up over 100 µV. A less mature neonate will have more weak signal parts and

more spikes with an amplitude above 100 μV.

5.1.1 Correlated features

The model included only 5 of the 12 investigated features. This was because many features had a high correlation to each other, and needed to be removed to produce reliable feature-importance results. The fact that they were removed does however not mean that they do not hold importance for the prediction. The results in table 4.2 show for example that the correlation is very high between the low envelope mean and the time above 10 μV as well as the high envelope mean, at 0.97 and 0.98 respectively. As these values are close to full correlation, they are almost completely interchangeable with the feature regarded as most important, meaning that they also hold high predictive power for the output compared to the other features. But if all of them were included in the model, then the feature importance methods would have considered all of them less important as the three features would share the importance approximately three ways. The point being that while all of them are good indicators of brain age, all of them do not need to be included in an explainable ML model because of their correlation; it is enough to include one of them and including more than one is excessive.

5.1.2 Limitations of Feature Importance Results

The main limitation of this result is that these feature-importance results are based on a model that can explain 45% of the variability of the data. That means that there is a relatively large amount of variability that is unexplained. This means that while these results provide some information on what is important for predicting maturation in a regression-tree model, there are possibly other signal features that also contribute to accurate predictions that are not investigated in this study. If the model would perform very well, then the feature-importance results would be of higher impact. In that case it would be possible to conclude that the result shows all the necessary features in order of importance for accurately predicting maturation for this type of model. This is however not the case.

5.1.3 Improving the Result

To increase the impact of the result it would be necessary to increase the performance of the model. This could be done by finding features that are both uncorrelated to the ones already investigated and to some degree correlated

with brain maturation. One way do to this would be to investigate more features in the frequency domain (which would be better to do on EEG and not aEEG, as the aEEG is rectified) or features that investigate the characteristics of the bursts in the signal and the inter-burst interval (which is not possible in 10-minute segments).

Another way to find new features would be to not decouple the right and left channels of the EEG and look at measurements that require two channels, like brain symmetry index and cross-correlation. If that would be done then the one-channel features in this work should only be calculated on one of the channels and not both, as they are highly correlated.

If these features would improve the performance of the model, then they would also improve the impact of the result of this study, as well as make the list of investigated features longer.

As this is a master's thesis, which infers a limited timespan, and as the data required processing which was more time-consuming than expected, the time that could be spent on extracting and investigating features was not as much as was hoped at the onset of the project. The original plan was to investigate a longer list of features than the one investigated.

5.2 Regression Model Performance

The regression model's R^2-score of 0.45 indicates that the model explains almost half of the variability in the data. The RMSE of the model is 1.73 weeks and its correlation coefficient between the output and the real labels is 0.676. This is clearly better than the chance level of 2.33 weeks, but not as good as from other works within the field, see table 5.1 for a comparison of this model with models from other works.

In the table it can be seen that this model is not as effective as models from other works. This could be explained by the following reasons:

- The model used in this work is explainable and is therefore likely to perform worse due to the performance-explainability tradeoff as mentioned in section 2.2.1.
- It is also only using 5 features compared to the other models which use more.
- The other studies use EEG signals and not aEEG, which makes it possible to extract more useful spectral features.
- The other studies have larger datasets.

Table 5.1: Results from this study and other studies. Some studies used other evaluation metrics than this study, therefore some values are missing.

Model	RMSE [weeks]	R^2	PCC	Model Type
Model in this project	1.73	0.45	0.676	Regression Tree
Chance model	2.33	0	-	Dummy model
Model from [5]	1.29	-	0.833	Support Vector Regression
Model from [6]	-	-	0.936	Support Vector Regression
Model from [37]	-	-	0.966	Gradient Boosted Model

Model	Features	Length of segment	Number of electrodes
Model in this project	5	10 minutes	4
Chance model	5	10 minutes	4
Model from [5]	41	1 hour	9
Model from [6]	46	1 hour	9
Model from [37]	59	Mean length 2.7 hours	9

- This model uses 10-minute segments while the other studies use 1-hour segments, 1-hour segments and a data set with an average recording time of 2.7 hours respectively.
- This model uses single-channel EEG data while the other models use multiple channels.

For comparison with clinical performance, visual assessment of aEEG by a group of internationally recognized experts with 7 or more years of experience was found to have a trend of underestimating the PMA by 1.8 weeks, as mentioned in subsection 2.4. So the model developed in this work is comparable to the performance of the experts. It should however be pointed out that the clinicians in [7] that assessed EEG performed better than the model in this study. Assessing EEG is however more time-consuming than assessing aEEG and is not always employed in NICUs.

As the model is possible to improve in multiple aspects without compromising the explainability, the outlook for the development of an ML model that can help clinicians in NICUs is positive.

5.3 Other Technical Considerations

5.3.1 Small Dataset

The dataset used in this project can be considered small, with signals collected from only 14 neonates. This limits the ability to draw strong conclusions from the result, and the result should be considered as an indication of which features are of importance and how well an explainable model can perform on

this type of data. A larger dataset collected from different clinics and a larger number of neonates would make it possible to draw stronger conclusions from the result and achieve better predictive results.

The dataset only consists of neonates with a PMA up to 32.1 weeks, spanning from extremely preterm up to very preterm. A dataset with neonates aged up to 37 weeks, the age at which a neonate is classified as preterm, would give a result applicable to the whole range from extremely preterm to late preterm.

5.3.2 10-Minute Segments

The fact that a 10-minute window was chosen makes the model treat periods with bursts the same as periods without bursts because a 10-minute window is normally not enough to capture a full period of bursts. As the EEG for neonates typically has these periods with and without bursts, some information is lost when treating all windows the same, and the best signal characteristic for the prediction might be specific to the behaviour of the periods with or without bursts. For instance, the mean value of the low envelope in a burst period might be a better predictor of maturation than just generally assessing the mean value of the low envelope. An alternative design choice could be to implement a burst detector that first finds out if the current segment is in a burst period or not and make predictions with that information in consideration.

5.4 Other Aspects of the Work

Following is a discussion of the work from other aspects than the purely technical, and putting it into a larger context. Ethical, economic, societal and sustainable aspects of the work will be discussed.

5.4.1 Ethical Aspects

The ethical discussion will follow the framework of principlism [43] by Beauchamp and Childress and its four principles: autonomy, beneficence, non-maleficence and justice.

The principle of autonomy of the patients can be considered non-applicable as the patients are not able to make their own decisions. It is however applied to the parents; the parents should be able to decide the care for their child. This also includes that the parents should be able to control what data is collected from their child and how it is used in the AI, and they should also have access

to understandable explanations of how the AI algorithms work and how it is used in their child's care. This emphasizes the importance of explainability in medical AI models. The parents have the right to know the reasoning behind a suggested care decision, and this includes if the decision was reached through AI algorithms.

The principle of beneficence states that healthcare providers should act in the best interest of their patients, and AI models here can help in making healthcare decisions faster and more accurate and develop more effective treatment plans. This project contributes in this regard by helping produce models which could help identify maturational deviations early in the pre-term neonate to facilitate necessary interventions.

The principle of non-maleficence states that no harm should be caused to patients. This emphasizes creating accurate, stable and trustworthy models. Working with EEG signals, this accents the need for effective artefact subtraction, disconnect detection and effective removal of parts of the signal that are faulty. The model must have proper ways to deal with faulty parts of the signal, as producing an incorrect estimation could increase the risk of the clinician making an incorrect decision and unknowingly causing maleficence for the patient. This is why this project implements three ways to remove or correct faulty parts of the signal (artefact subtraction, thresholding in the time domain and removal of datapoints with unrealistic feature values). Algorithms should be designed in a way to minimize the risk of harm to patients.

The principle of justice refers to the fair and equitable distribution of healthcare resources. In the context of this project it implies that the data used to train the models should be diverse and representative of the patient population to avoid possible biases, such as gender or racial bias, that makes the model more efficient for some groups of patients than other groups. This can result in unfair treatment of certain groups of people. This is also a reason why explainability is important, as it is possible to inspect how decisions were made, and it facilitates investigating and adjustment of a model that might turn out to be unjust. This is more difficult in black box models.

5.4.2 Economical, Social and Sustainability Aspects

Economical Aspects

Economically, this work contributes to simplifying medical measurements and health assessments. This means that less of the clinicians' time and energy needs to be spent on making this assessment. This saves costs as it is reducing the clinicians' mental workload and a clinician will be able to make better

decisions, thus reducing costs from bad decisions, as well as being able to take care of more patients.

Reducing costs in healthcare makes care available to more people. This also follows the ethical principle of justice; more people can get the care they need, and fewer decisions need to be taken regarding who receives care or not.

Social Aspects

This work is a contribution towards developing ML models that can be accepted by the medical community and practically implemented in NICUs, by virtue of their explainability. The specific results of this study are of interest to those who aim to develop these types of models. Developing a reliable method for automatically assessing brain maturation is of interest to neonatologists as it makes the task quicker and less demanding, the hospital as it makes the care less costly, and the parents and the neonates as the neonate receive better care. The fact that streamlining medical measurements reduces costs will also contribute to making this type of care available to more people. In a larger context, it is also of interest to society as decreasing neonatal death and injury contributes to a healthier population.

Sustainability

The work and its results do not have a strong connection to sustainability. The results at this stage have no impact on the environment as it is just contributing knowledge. But supposing a created final model that can be distributed that is based on this work: It does not change the way the measurement is taken, thus neither reducing nor increasing the number of resources and materials that the hospital uses. While the training of an ML model might need a powerful computer, using the final model is less computationally heavy and the hospitals should be able to use the computers they already have and not need to buy new ones to use the model, although this does depend on how complex the final model would be. In the case that every NICU that the model is implemented in needs to upgrade their computers, it will have an impact on the environment in the form of increased resource consumption.

Chapter 6

Conclusions and Future work

The conclusions from this study are presented and followed by a section regarding future work.

6.1 Conclusions

An explainable machine learning model was developed for estimating post-menstrual age in very preterm neonates based on EEG signals. The model produced estimations with a root mean square error of 1.73 weeks. The most important signal feature for the model was the mean value of the low envelope followed by signal time over 100 µV, but other features were found to have a high correlation with these and are also of importance when predicting age. The more premature the neonate is, the more weak parts (low amplitude) exist in the signal, as well as more spikes to amplitudes above 100 µV compared to a less premature neonate. The model's performance can be improved and given that it performs comparably to human experts, these results indicate a promising outlook for explainable machine learning in neonatal EEG analysis. The findings are an important first step towards developing explainable machine learning models that can be implemented in neonatal intensive care units to aid clinicians.

6.2 Future Work

This work is a first step towards an ML model that can be practically used in NICUs thanks to its explainability, but more work is needed in this field for explainable ML to be used in practice. A logical continuation of this work is to develop a tree-based ML model with a focus on improving the performance

of the model's estimation of neonatal brain maturation from aEEG or EEG data as the performance has to be improved to be practically used in NICUs. The improvements can be made by including more features in the model, using EEG data instead of aEEG and having a larger dataset.

Another continuation would be to develop a functional explainable ML model and introduce it to clinicians and get feedback from them on what they think about the model and how they would use it in practice and ask them if the level of explainability is satisfactory in clinical practice. Also asking them to evaluate general impressions and user-friendliness would be of interest to make a model easily adoptable in NICU.

A model could also be designed to assess other information than brain maturation. It could for instance assess pathological deviations and specific neurological conditions. The model could also be expanded with labelled input data, which for instance could inform the model if a neonate is on a specific medication, and therefore predict the maturation and other outcomes if the model has been trained on other neonates on that medication with known outcomes. In short, it should be possible to develop a user-friendly program based on explainable machine learning to be used in clinics that assesses more aspects of the neonates' neurological health than only maturation.

References

[1] J. E. Lawn *et al.*, "Small babies, big risks: global estimates of prevalence and mortality for vulnerable newborns to accelerate change and improve counting," *The Lancet*, may 2023. doi: 10.1016/s0140-6736(23)00522-6

[2] J. Perin, A. Mulick, D. Yeung, F. Villavicencio, G. Lopez, K. L. Strong, D. Prieto-Merino, S. Cousens, R. E. Black, and L. Liu, "Global, regional, and national causes of under-5 mortality in 2000–19: an updated systematic analysis with implications for the sustainable development goals," *The Lancet Child & Adolescent Health*, vol. 6, no. 2, pp. 106–115, feb 2022. doi: 10.1016/s2352-4642(21)00311-4

[3] A. W. Gill, "Postnatal cardiovascular adaptation," *Archives of Disease in Childhood - Fetal and Neonatal Edition*, vol. 104, no. 2, pp. F220–F224, jul 2018. doi: 10.1136/archdischild-2017-314453

[4] J. O'Toole and G. Boylan, "Quantitative preterm EEG analysis: The need for caution in using modern data science techniques," *Frontiers in Pediatrics*, vol. 7, may 2019. doi: 10.3389/fped.2019.00174

[5] J. O'Toole, G. Boylan, S. Vanhatalo, and N. Stevenson, "Estimating functional brain maturity in very and extremely preterm neonates using automated analysis of the electroencephalogram," *Clinical Neurophysiology*, vol. 127, no. 8, pp. 2910–2918, aug 2016. doi: 10.1016/j.clinph.2016.02.024

[6] N. J. Stevenson, L. Oberdorfer, N. Koolen, J. M. O'Toole, T. Werther, K. Klebermass-Schrehof, and S. Vanhatalo, "Functional maturation in preterm infants measured by serial recording of cortical activity," *Scientific Reports*, vol. 7, no. 1, oct 2017. doi: 10.1038/s41598-017-13537-3

[7] N. J. Stevenson, M.-L. Tataranno, A. Kaminska, E. Pavlidis, R. R. Clancy, E. Griesmaier, J. A. Roberts, K. Klebermass-Schrehof, and

S. Vanhatalo, "Reliability and accuracy of EEG interpretation for estimating age in preterm infants," *Annals of Clinical and Translational Neurology*, vol. 7, no. 9, pp. 1564–1573, aug 2020. doi: 10.1002/acn3.51132

[8] A. Okumura, F. Hayakawa, T. Kato, K. Watanabe, and K. Kuno, "Developmental outcome and types of chronic-stage EEG abnormalities in preterm infants," *Developmental Medicine &amp Child Neurology*, vol. 44, no. 11, pp. 729–734, feb 2007. doi: 10.1111/j.1469-8749.2002.tb00278.x

[9] D. Bonthius, "Subacute sclerosing panencephalitis, a measles complication, in an internationally adopted child," *Emerging Infectious Diseases*, vol. 6, no. 4, pp. 377–381, aug 2000. doi: 10.3201/eid0604.000409

[10] B. C. Oxley, "International 10-20 system for eeg electrode placement, showing modified combinatorial nomenclature," Online, Jul. 2017. [Online]. Available: https://commons.wikimedia.org/wiki/File:International_10-20_system_for_EEG-MCN.svg

[11] M. Cordeiro, H. Peinado, M. T. Montes, and E. Valverde, "Evaluation of the suitability and clinical applicability of different electrodes for aEEG/cEEG monitoring in the extremely premature infant," *Anales de Pediatría (English Edition)*, vol. 95, no. 6, pp. 423–430, dec 2021. doi: 10.1016/j.anpede.2020.09.010

[12] Z. A. Vesoulis, P. G. Gamble, S. Jain, N. M. E. Ters, S. M. Liao, and A. M. Mathur, "WU-NEAT: A clinically validated, open-source MATLAB toolbox for limited-channel neonatal EEG analysis," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105716, nov 2020. doi: 10.1016/j.cmpb.2020.105716

[13] K. T. Tapani, P. Nevalainen, S. Vanhatalo, and N. J. Stevenson, "Validating an SVM-based neonatal seizure detection algorithm for generalizability, non-inferiority and clinical efficacy," *Computers in Biology and Medicine*, vol. 145, p. 105399, jun 2022. doi: 10.1016/j.compbiomed.2022.105399

[14] S. Knapič, A. Malhi, R. Saluja, and K. Främling, "Explainable artificial intelligence for human decision support system in the medical domain," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 740–770, sep 2021. doi: 10.3390/make3030037

[15] A. Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer, 2017.

[16] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?" 2017.

[17] F. K. Dosilovic, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: A survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, may 2018. doi: 10.23919/mipro.2018.8400040

[18] D. Raab, A. Theissler, and M. Spiliopoulou, "XAI4eeg: spectral and spatio-temporal explanation of deep learning-based seizure detection in EEG time series," *Neural Computing and Applications*, sep 2022. doi: 10.1007/s00521-022-07809-x

[19] J. Amann, , A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, nov 2020. doi: 10.1186/s12911-020-01332-6

[20] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, vol. 8, pp. 42 200–42 216, 2020. doi: 10.1109/access.2020.2976199

[21] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, no. 1, oct 2019. doi: 10.1186/s12916-019-1426-2

[22] S. Kundu, "AI in medicine must be explainable," *Nature Medicine*, vol. 27, no. 8, pp. 1328–1328, jul 2021. doi: 10.1038/s41591-021-01461-z

[23] D. F. Hamilton, M. Ghert, and A. H. R. W. Simpson, "Interpreting regression models in clinical outcome studies," *Bone &amp Joint Research*, vol. 4, no. 9, pp. 152–153, sep 2015. doi: 10.1302/2046-3758.49.2000571

[24] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[25] C. Molnar, *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable*, 2nd ed., 2022, chapter 8.5.

[26] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017.

[27] L. S. Shapley, "17. a value for n-person games," in *Contributions to the Theory of Games (AM-28), Volume II*. Princeton University Press, dec 1953, pp. 307–318.

[28] S. Lundberg, G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," Feb. 2018.

[29] E. L. Lydia, C. S. S. Anupama, and N. Sharmili, "Modeling of explainable artificial intelligence with correlation-based feature selection approach for biomedical data analysis," in *Biomedical Data Analysis and Processing Using Explainable (XAI) and Responsive Artificial Intelligence (RAI)*. Springer Singapore, 2022, pp. 17–32.

[30] V. F. Burdjalov, S. Baumgart, and A. R. Spitzer, "Cerebral function monitoring: A new scoring system for the evaluation of brain maturation in neonates," *Pediatrics*, vol. 112, no. 4, pp. 855–861, oct 2003. doi: 10.1542/peds.112.4.855

[31] S. Kesić and S. Z. Spasić, "Application of higuchi's fractal dimension from basic to clinical neurophysiology: A review," *Computer Methods and Programs in Biomedicine*, vol. 133, pp. 55–70, sep 2016. doi: 10.1016/j.cmpb.2016.05.014

[32] R. Esteller, G. Vachtsevanos, J. Echauz, and B. Litt, "A comparison of waveform fractal dimension algorithms," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 48, no. 2, pp. 177–183, 2001. doi: 10.1109/81.904882

[33] B. Koley and D. Dey, "An ensemble system for automatic sleep stage classification using single channel EEG signal," *Computers in Biology and Medicine*, vol. 42, no. 12, pp. 1186–1195, dec 2012. doi: 10.1016/j.compbiomed.2012.09.012

[34] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Physica D: Nonlinear Phenomena*, vol. 31, no. 2, pp. 277–283, jun 1988. doi: 10.1016/0167-2789(88)90081-4

[35] T. Inouye, K. Shinosaki, H. Sakamoto, S. Toi, S. Ukai, A. Iyama, Y. Katsuda, and M. Hirano, "Quantification of EEG irregularity by use of the entropy of the power spectrum," *Electroencephalography and Clinical Neurophysiology*, vol. 79, no. 3, pp. 204–210, sep 1991. doi: 10.1016/0013-4694(91)90138-t

[36] "Detection, classification, and estimation in the (t , f) domain," in *Time-Frequency Signal Analysis and Processing*. Elsevier, 2016, pp. 693–743.

[37] X. Dong, Y. Kong, Y. Xu, Y. Zhou, X. Wang, T. Xiao, B. Chen, Y. Lu, G. Cheng, and W. Zhou, "Development and validation of auto-neo-electroencephalography (EEG) to estimate brain age and predict report conclusion for electroencephalography monitoring data in neonatal intensive care units," *Annals of Translational Medicine*, vol. 9, no. 16, pp. 1290–1290, aug 2021. doi: 10.21037/atm-21-1564

[38] M. Waskom, "seaborn: statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, apr 2021. doi: 10.21105/joss.03021

[39] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, sep 2020. doi: 10.1038/s41586-020-2649-2

[40] W. McKinney *et al.*, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, vol. 445, 2010, pp. 51–56.

[41] A. S. Morgan, M. Mendonça, N. Thiele, and A. L. David, "Management and outcomes of extreme preterm birth," *BMJ*, p. e055924, jan 2022. doi: 10.1136/bmj-2021-055924

[42] M. V. Aparicio, "Meeting notes," Personal communication, Mar. 2023.

[43] T. Beauchamp and J. Childress, *Principles of Biomedical Ethics*. Oxford University Press, 1994. ISBN 9780195085365

Appendix A

Supplementary Data

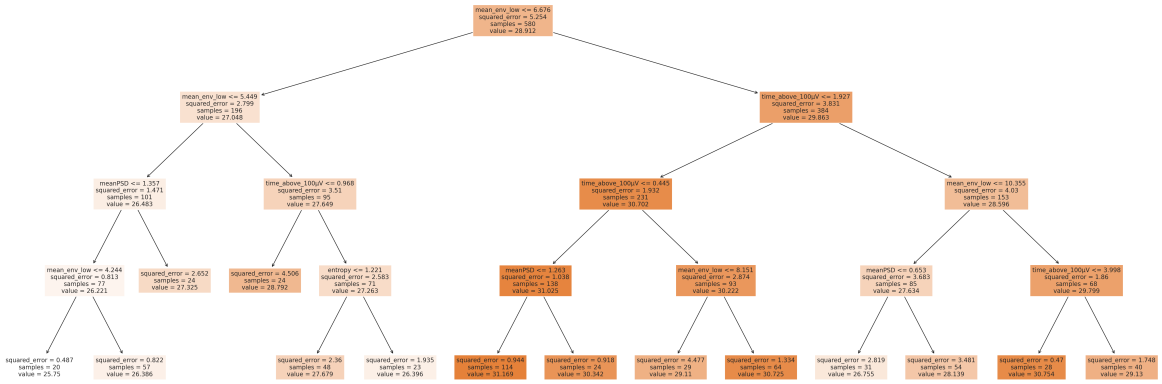

Figure A.1: Overview of the full decision tree. The darker orange a node is, the higher the estimated age is. For a detailed view, see figure A.2.

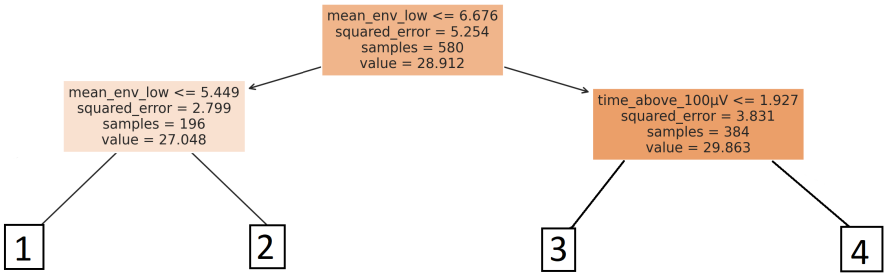

Figure A.2: Detailed view of the top nodes in the model. See figure A.3 and A.4 for the branches 1 through 4.

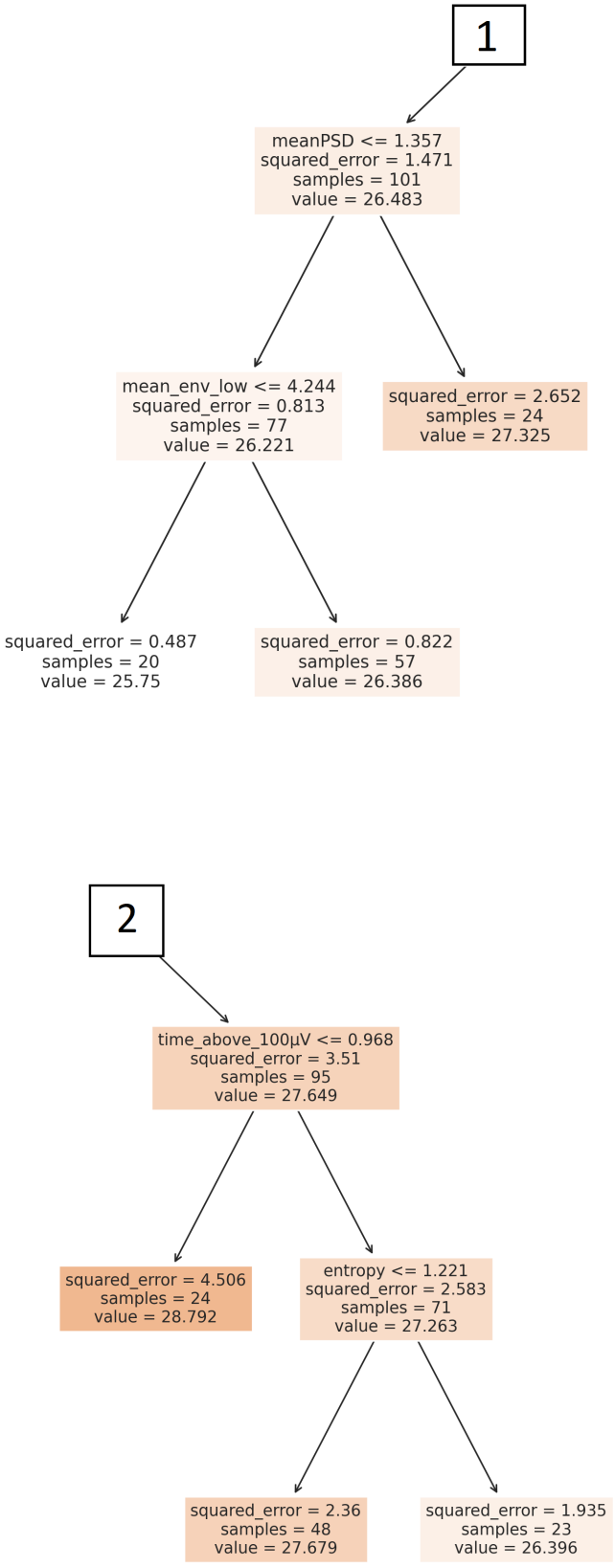

Figure A.3: Detailed view of branches 1 and 2.

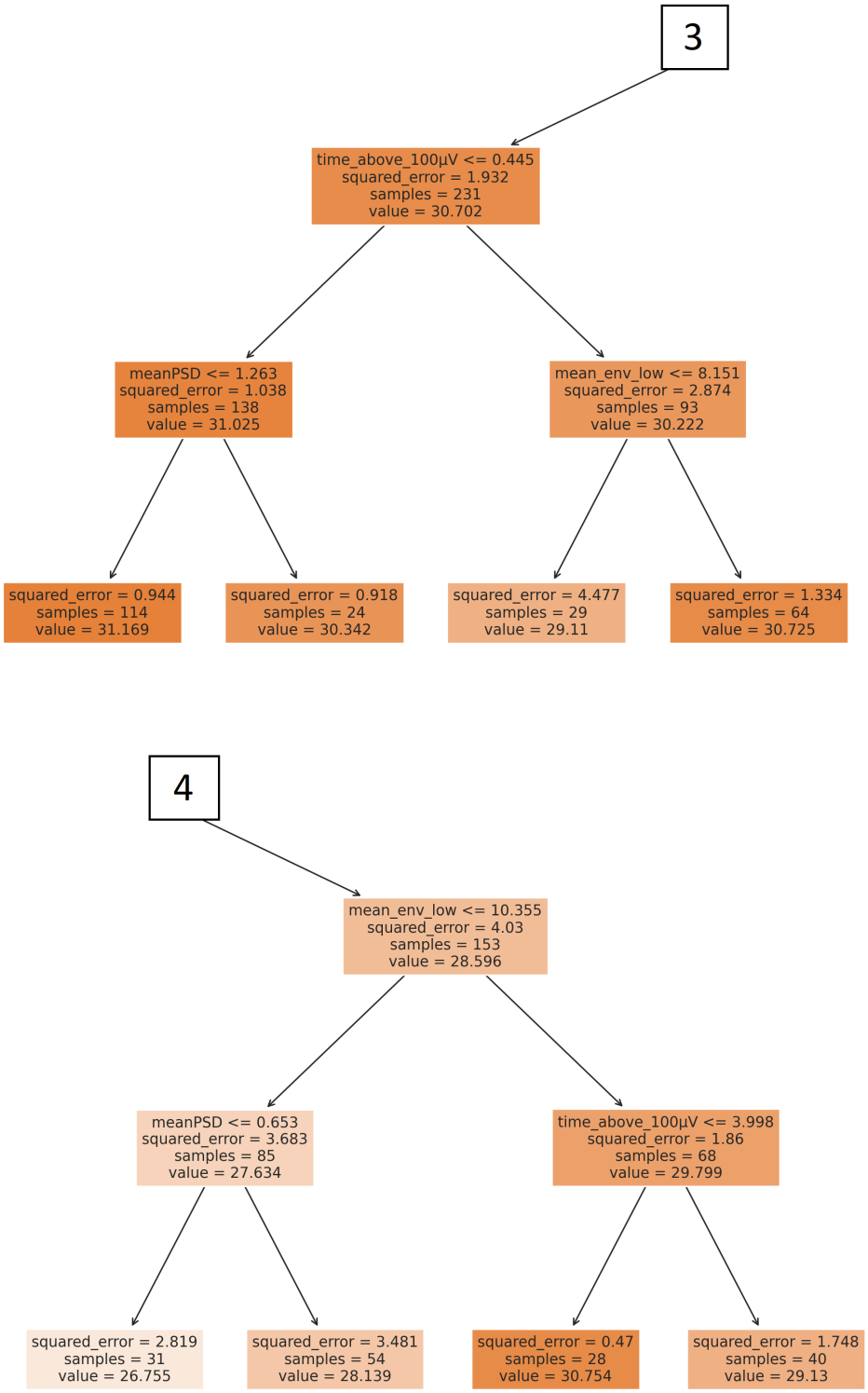

Figure A.4: Detailed view of branches 3 and 4.

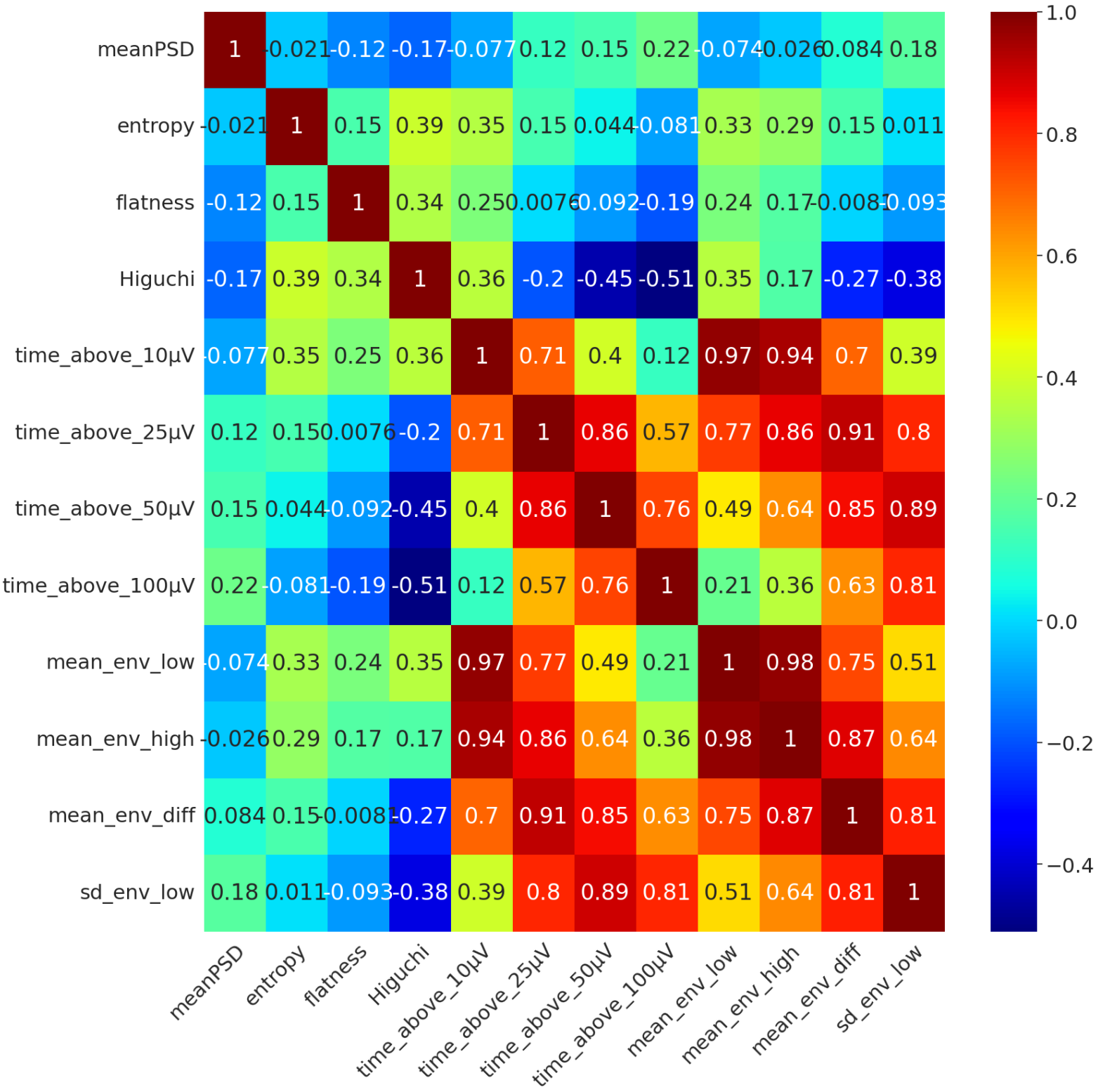

Figure A.5: The full correlation matrix with correlation values

For DIVA


```
{
"Author1": {
            "Last name": "Svensson",
            "First name": "Patrik",
            "E-mail": "patrsve@kth.se",
            "organisation": {"L1": "School of Electrical Engineering and Computer Science ",
                                }
          },
"Degree": {"Educational program": "Master's Programme, Medical Engineering, 120 credits"},
"Title": {
            "Main title": "Estimating Brain Maturation in Very Preterm Neonates",
            "Subtitle": "An Explainable Machine Learning Approach",
            "Language": "eng" },
"Alternative title": {
            "Main title": "Estimering av hjärnmognad i mycket prematura spädbarn",
            "Subtitle": "En ansats att tillämpa förklarbar maskininlärning",
            "Language": "swe"
},
"Supervisor1": {
                "Last name": "Benítez",
                "First name": "Raúl",
                "E-mail": "raul.benitez@upc.edu",
                "Other organisation": "Universitat Politècnica de Catalunya"}
              },
"Examiner1": {
                "Last name": "Larsson",
                "First name": "Matilda",
                "E-mail": "matil@kth.se",
                "organisation": {"L1": "School of Engineering Sciences in Chemistry, Biotechnology and Health ",
}
              },
"Other information": {
"Year": "2023", "Number of pages": "xi,56"}
}
```

TRITA – CBH-GRU 2023:149