ANALYSIS OF MACHINE LEARNING MODELS

ON INFANT MOVEMENT DATA

ANALYSIS OF MACHINE LEARNING MODELS ON INFANT MOVEMENT DATA

By OMAR NASSIF, B.Sc.,

A Thesis Submitted to the School of Graduate Studies in the Partial Fulfillment of the Requirements for the Degree Master of Applied Science

McMaster University

Master of Applied Science (2020)

Hamilton, Ontario (Department of Electrical and Computer Engineering)

TITLE: Analysis of Machine Learning Models on Infant Movement Data

AUTHOR: Omar Nassif

B.Sc., (Electrical Engineering), McMaster University,

Hamilton, Ontario, Canada

SUPERVISORS: Drs. James P. Reilly, Victoria Galea

NUMBER OF PAGES: xi, 66

Abstract

This thesis presents a study of feature engineering and supervised models on infant general movements. General movements are purposeless movements produced by infants that can be used by clinicians to evaluate an infant's developmental health. Given a database of healthy infant movement recordings, we train both supervised models and clustering algorithms to gain clinical insight into the data. First, a large set of time domain and frequency domain features are calculated to extract clinically meaningful features from the raw movement data. The infants' data were split into different age groups based on the range of age, with the age group being used as the training label. Then various supervised models were trained on age group labels to predict the age group of an infant given their movement features. Using appropriate validation schemes, the supervised models attained good sensitivity and specificity on out-of-sample subjects, but also reflected large physiological variance by showing overlap between the age groups. Finally various future directions are presented most important of which is applying clustering algorithms with some preliminary results showing interesting clusters of infants. Overall these results show how healthy infants in the early months of life move, and are conducive to further studies that can quantify at-risk infant development relative to healthy infants.

Acknowledgements

I would like to sincerely thank my supervisors, Dr. James Reilly and Dr. Victoria Galea. They have guided me on this journey through my graduate studies on all aspects of it. I have learned a tremendous amount from them on how to conduct research and was given many opportunities to collaborate with colleagues through them. I want to thank the MacDATA Institute for sponsoring part of my research. I learned a lot about various applications of machine learning to the real world and this has opened my eyes beyond my research project. I also want to thank Dr. Simon Overduin for useful discussions related to the processing of movement data. Last but not least I want to thank my family and my especially parents for their continued support throughout my graduate studies, taking an interest in a subject they know nothing about just to lend me an ear.

Table of Contents

List of Figures

List of Tables

List of Acronyms

CP	Cerebral Palsy
EM	Expectation Maximization
GM	General Movement
GMM	Gaussian Mixture Model
LkSO	Leave-k-Subjects-Out
MFA	Mixture of Factor Analyzers
MAD	Mean Absolute Deviation
ML	Machine Learning
RF	Random Forest
SAL	Spectral Arc Length
SNR	Signal-to-Noise Ratio
STD	Standard Deviation
SVM	Support Vector Machine

Chapter 1

Introduction

1.1 Background

General movements (GMs) are defined as purposeless movements that infants produce using all their limbs and with no particular pattern (Hadders-Algra., 2004). These movements can be observed at birth and throughout the first weeks of life up to around 4 months of age, when a baby is laid on their back with their limbs unconstrained. GMs are produced from the time of birth up to around the age of 5 months, after which movements start becoming goal-oriented. Despite the fact that GMs are seemingly random, close evaluation of GMs by trained clinicians can lead to insight into the infant's brain development and health (Einspieler and Prechtl 2005; Hadders-Algra., 2007). A connection was established linking neural development with spontaneous movement outbursts and so GMs can be thought of as a window into the real-time development of the infant's nervous system (Feller., 1999). This led to the conclusion that the evaluation of GMs can be a

powerful, non-invasive, and quick diagnostic tool of the developing nervous system (Prechtl,. 1990).

A particular application of GM diagnosis is the ability to, from very early on in the child's life, predict later developmental disorders (Prechtl., 1990). Specifically GM diagnoses has been applied to the prediction of cerebral palsy (CP) and a strong correlation has been established to show abnormal GMs predict cerebral palsy with high sensitivity and specificity (Prechtl et al., 1997; Bosanquet et al., 2013). GM diagnosis is also being applied for prediction of developmental cognitive delays (Kanemaru et al., 2013).

The diagnosis of cerebral palsy and other developmental disorders is generally difficult to establish from an early point in a child's life. Traditional diagnosis are established around 2-5 years of age depending on the condition and severity level (Cans., 2000). Late diagnoses miss a crucial window of time in which interventional therapies should have been applied. It has been shown that the earlier the therapy, the better the outcome (Liang et al., 2014; Hebbeler et al., 2007). On the other hand, GM diagnosis can be applied very early on to at-risk infants at ages 1-4 months. The benefit of this is of course to give an indicator of whether or not the infant will require interventional therapy. Such early intervention has been shown to yield positive results in comparison to developmentally delayed children who have not had a specialized intervention (Blauw-Hospers et al., 2007). However despite the benefits, the diagnosis of GMs requires a clinician to have many years of training, and the diagnosis can be subjective. Currently GM diagnosis are carried out by taking a video of a baby and then sending it to a GM clinic where trained experts will diagnose the infant

in the video. In order to ensure wider implementation of this diagnosis, an automated diagnosis via a machine learning approach is a promising route to take.

As with any machine learning model, we have to understand the data and what features should be looked at that are also clinically relevant. There is vast literature about the qualities and classification of GMs that are of interest to a clinician. GM patterns vary from infant to infant, however, qualitatively there are specific ways to categorize GMs. Some repeated keywords across the literature are the complexity, variability, and fluency of the GMs (Taga et al., 1999; Hadders-Algra., 2014). As Hadders-Algra describes in her 2014 paper, complexity is "the spatial variation of the movements", variability is "the temporal variation of the movements. It means that across time, the infant produces continuously new movement patterns" and fluency is essentially a measure of smoothness of translations and rotations while changing direction. As infants age in the first 3 months, their movement complexity, fluency, and variability, rapidly develop.

A machine learning model capable of classifying GMs according to these named qualities would be clinically useful due to its interpretability and portability. Not only that, but analyzing the GMs with machine learning tools that extract clinically relevant features may lead to new physiological insight and reveal subtle movement signatures that were previously unnoticeable by a human observer.

1.2 Problem Statement

In this thesis however, we focus on the slightly different problem, of classifying age groups given movement data features. Typically pathological movement patterns show that older infants are lagging behind their general age group, and so by building models that can correctly classify or group healthy infants, we can measure deviation from this "healthy" developmental trajectory. Further, we can use such models to isolate which features of movement are most discriminative of healthy development.

Studies show that the qualitative aspects of GMs change rapidly during the first few months of life and there are transitional periods where things generally change. Einspieler and Prechtl (2005) review the taxonomy of GMs for both normal and abnormal GMs as:

Age Group	Type of GM
< 8 Weeks Old	Writhing movements: Typically elliptical movement of the limbs.
> 8 Weeks Old	Fidgety movements: small circular movements of small amplitudes of the limbs. Increased movement fluency.

Table 1.1 Taxonomy of GMs for each approximate age group (Einspieler and Prechtl 2005)

As with most human related experiments, gathering large quantities of good data is difficult. This difficulty is further increased when dealing with infants as young as 2 – 16 weeks old. In this problem setting, we have gathered a database of general movements recorded from

49 healthy infants and therefore we focus solely on healthy infant movement classification and clustering. The task is to train various machine learning models on movement features that will allow us to quantify how healthy infants move with respect to an age group. Our hypothesis is that, different age groups have different defining characteristics and by using movement features we can predict which age group an infant falls into with significant accuracy.

However, a priori we also expect there to be a lot of overlap between the different age groups due to the large variance that naturally exists in physiological data. And so in reality the problem becomes less of a strict classification problem, and more of a clustering problem with noisy labels, which is also referred to as semi-supervised learning. Further, not all features will show the general trends that exist in each age group. Naturally some features will not discriminate between major age group movement trends and will only contribute noise, and therefore confound any interesting results. And so along with a clustering task we will also need a feature selection or dimensionality reduction that can highlight which feature set best shows clear clustering of different age group movement characteristics. As discussed in the introduction, clinicians typically score three measures or factors of infant movements, which are complexity, fluency, and variability. A useful feature reduction algorithm should project our feature space on these three factors to give us better interpretability.

To reiterate, the tasks are three fold:

- Feature engineering of interpretable features that reflect aspects of complexity, fluency, and variability of GMs
- Supervised classification of infant movement features into the given age group with different levels of detail in age prediction

1.3 Summary of Thesis Contributions

This thesis studies the applicability of supervised and semi-supervised models on infant movement data. Here we point out the major contributions of this thesis.

1.3.1 Movement data features

As interpretability is a big focus for useful clinical models, interpretable features are calculated and extracted from the infant kinematic data. The features were chosen to reflect some aspect of complexity, fluency, and variability, which are standard clinical measures. Some novel features are proposed that give the classifiers higher accuracy than more traditional choices of features such as velocity statistics.

1.3.2 Supervised classifiers

Two standard classifiers are trained and tuned with a big focus on accurate estimation of generalization error. The first is a support vector machine (SVM) model with a linear kernel that linearly separates the data demonstrating that the classes are linearly separable to a significant degree, with some overlap existing between classes in feature space. The second classifier trained is a random forest (RF) classifier that has a non-linear decision boundary. The SVM and RF classifiers both showed good accuracy but with different sensitivities and specificity on the class errors. The classifiers also give feature weights that showed which features best affected the classifiers' decision to discern between age groups.

1.3.3 Discussion of future directions

A Mixture of Factor Analyzers (MFA) model, was used to apply simultaneous dimensionality reduction and clustering with only 30% of labels provided. 70% of age group labels were purposely withheld from the algorithm to cluster the data because of the displayed overlap with the classifiers of the previous section. Though there is a big caveat in that our data samples multiple points from the same infant which skews statistical models such as the mixture models used here. Nevertheless some interesting clustering is shown that indicates that more sophisticated clustering algorithms can reveal age related structure within the data.

1.4 Organization of the Thesis

First the experimental design is discussed to ground the thesis data and analysis, such as what the data is and how it was acquired, processed, and segmented into the final data table. Then a literature review of classification of infant kinematics is presented that involves various applications such as diagnosing pathologies and classification of age groups, the latter being the specific topic of this thesis. Then features of movement data are discussed along with their modelling capability of movement complexity, fluency, and variability. After that, the classification section is presented and finally the conclusion and future directions of this research is discussed briefly.

Chapter 2

Literature Review

In this section we go over some of the literature that utilizes machine learning on various tasks that involve infant kinematics. Though only one of the studies deals directly with the specific task of age group separation, we use the other studies to build an intuition of which features are appropriate and what aspects of movements are interesting.

2.1 Classification of Infant Kinematics

2.1.1 Classifying infants based on abnormal GMs

In 2015 a paper was published reviewing various analysis techniques being applied to GM evaluation (Marcroft et al., 2015). This paper showed the approach of 12 groups from around the world. The various approaches were categorized according to the data. Some groups extracted GM features from video data by looking at things like optical flow (indirect sensing) whereas others used some form of motion tracking devices to directly measure the GMs (accelerometers, magnetic tracking system, and etc.). The groups with

direct measurements are of particular interest to this thesis due to the similarity in recorded data type.

Gravem et al. (2012) had data acquired from an accelerometer. Their data was acquired from 10 infants, 6 who were later diagnosed with cerebral palsy and 4 healthy babies. They extracted statistical features like mean, standard deviation, min, and max of the accelerometer recorded values. They used a combination of SVM and DBN to classify their results. While they were able to see some form of clustering between healthy and pathological data, due to their small sample size it is difficult to generalize the results. Kanemaru et al. (2013) had a significantly larger sample size of 145 infants of which 16 were later diagnosed with cerebral palsy and 129 healthy infants. Their data were recorded from reflective markers placed on a baby and position extracted from a video recording. The main features for their analysis were what they termed "jerk index" (3^{rd} derivative of position), the kurtosis and skewness of acceleration, and correlation between limb velocities (Kanemaru et al., 2013). The jerk index is a measure of how smoothly the baby accelerated. Kurtosis and skewness of acceleration help to characterize the smoothness of the acceleration of the movements. In Karch et al. (2012) in which they come up with a mathematical model to measure "stereotypy". Stereotypy is defined as the repetition of movement patterns (Karch et al., 2012). This is similar to the correlation measure used by Kanemaru et al. (2013) however they measured a cross-correlation of the entire time recording. The model used by Karch et al. (2012) employs a Dynamic Time Warping approach where they compare movement segments against one another. Their results

indicated that with at-risk infants, their stereotypy scores are higher than the apparently healthy infants.

The findings of these are papers are very much in agreement with the qualitative descriptions established about normal and abnormal GM movements shown by Prechtl (1990) and others (Disselhorst-Klug et al., 2012). These findings also agree with some aspects of what Halders-Algra (2014) discussed about complexity, variation, and fluency of GMs, as explained in section 1.1 in the thesis.

2.1.2 Classifying infants based on age groups

The above studies dealt with the task of classifying GMs into healthy and pathological movements and they showed through features such as limb movement correlation and acceleration statistics this can be done. However the focus of this thesis is on classification into different age groups. The greater goal is gaining insight into how do GMs change over the first few months of life for a not-at-risk infant (i.e. healthy infant). Being able to correctly classify the approximate age group an infant falls into, leads us to identify which movement features best discriminate between the age groups and thus gain insight into the healthy developmental trajectory based on the discriminative features.

In Kato et al. (2014), they investigate environment-based behavior changes for GMs of infants who were in a slightly older age group range than ours. Their task was to determine if the older infants (~4 months of age) produced more complex movements compared with the younger infants (~3 months of age) during a play period with a toy. Data was recorded before play and during play. Arm and leg movements were recorded which were converted

into movement features such as statistics of amplitude and direction of velocity, as well as entropy of position. Their findings showed that the younger infants increased the amount of movement during play, but did not change their movement patterns. The older infants first changed their movement patterns during the play period and then increased the amount of movement. While the paper showed that there was no significant difference in movement patterns based on the entropy of position, their study differs from ours in that the age range was above ours, and in that they studied purposeful movements by introducing a toy into the experiment while our study focuses on purposeless movements.

Finally Disselhorst-Klug et al. (2012) attempt to directly model aspects of GMs mathematically that correspond to qualitative features that are manually assessed by clinicians. They focus on normal babies in the same age range as us (2 weeks old to 16 weeks old), and measure feature changes across age finding that many features were reflective of the infant being in a different developmental stage in early life with statistical significance. Though they note in their paper that none of the introduced features were able to express movement complexity and variability, nevertheless their features did show changes across the first few months of life for their healthy subjects.

The paper by Disselhorst-Klug et al. (2012) is the closest paper in terms of objective as it attempts to model qualitative aspects of GMs mathematically and to show how the changes of these feature measurements correlates with clinical knowledge of how GMs evolve as the infant ages. However, as the authors state, more sophisticated features were required that encompass complexity, variability and fluency of GMs in order to get the full picture. Their analysis was also limited to running an ANOVA on the different age groups to check

for statistical significance, however, in this thesis we will take this a step further to try to predict age group from the movement feature values themselves.

2.2 Commonly Used Features

Going over the papers above, we compile a list of features that are potentially useful for age group discrimination:

Feature	Feature Description
Statistics of Velocity	Mean, STD, Min, Max, Skewness of Velocity of arms and legs individually
Statistics of Acceleration	Mean, STD, Min, Max, Skewness of Acceleration of arms and legs individually
Stereotypy Score	Dynamic Time Warp to measure similarity of limb movements to one another
Positional Entropy	Dispersion of the probability of finding the limb in a some defined location during recording
Jerk	Jerk is used to model 'movement fluency' by taking the time integral of the 3rd derivative of movement normalized by movement duration
Cross-correlation between limbs	Correlation between arms and legs with all combinations in between

Table 2.1 Time-domain Features of Infant Movements

Most studies that looked at classification of normal and abnormal GMs looked at acceleration statistics (Marcroft et al., 2015). However, these feature values can be easily influenced by infant size if not normalized for correctly, and as we are looking to classify age, features which are affected by infant size are trivially correlated with age group. Further, for studies which measure position, calculation of the 2^{nd} and 3^{rd} numerical derivative introduces a lot of noise into the data significantly decreasing the SNR. In this thesis, we focus on features that are reflective of qualitative aspects of GMs that are not influenced by infant size and thus removing it as a confounding factor.

In the next section, the design experiment is discussed along with which features were selected that are reflective of complexity, fluency and variability of GMs.

Chapter 3

Experimental Design

In this section we discuss the experimental design, what data we are measuring, how it was acquired and the steps taken to process the raw data into a feature table that can be easily passed into a classification algorithm.

3.1 Data Acquisition

3.1.1 Modality and procedure

To acquire a database of healthy infant general movements, infants were recruited and their GMs were recorded using a device called the Trakstar Bird. This device sets up an electromagnetic field that can track (x, y, z) co-ordinates of sensors within the electromagnetic field in real-time, and then store these values for later processing. Because GMs are purposeless random movements, the experimental setup was as follows:

1. Infants were awake and as comfortable as possible with no stimulations present
2. Infants were placed on their back on a comfortable mat (Figure 3.1)

3. A Trakstar sensor is attached to each limb (right/left arm, right/left leg; Table 3.1)
4. With the infants laying on their back, their natural movements (GMs) were recorded for 3 – 5 minutes, and this is one trial
5. Trials were halted if an infant became too fussy, cried, or rolled off their back
6. Depending on the infant's emotional state and level of cooperation, two or even three trials were recorded
7. Infants were tested once during the first 8 weeks of life and, where possible, were asked to return for another session within their second 8 weeks. Usually around the 12th week of life

Figure 3.1 Experimental setup of infant GM data collection. The infant is laid on their back and allowed to freely move their limbs while their movements are recorded. The white tapes on the legs and arms are the sensors attached.

For each infant, multiple trials of 3 – 5 minute recordings each were recorded where the infant was sat on its back, and without any stimulation the natural movement of their limbs which we have referred to as general movements is recorded.

For each trial recording, we obtain a comma separated file (CSV) that contains x, y, z coordinates of 6 sensors. The Trakstar sensor attachment was as follows:

Sensor 1	Head
Sensor 2	Torso
Sensor 3	Right Arm
Sensor 4	Left Arm
Sensor 5	Right Leg
Sensor 6	Left Leg

Table 3.1 Sensor placement where each sensor records x, y, and z at 100Hz

Note that sensor 1 and sensor 2 were not used in any of the analyses.

3.1.2 User interface

For the Trakstar device, a user interface (Figure 3.2) that interfaced directly with the Trakstar device plotted sensor positions in real-time was programmed in C# and C++ to allow direct feedback ensuring that sensors were still within high quality range. Outside of this range sensor noise increased significantly and so this allowed us to ensure the data we obtained was accurate data.

Figure 3.2 The Trakstar user interface which was programmed and used for data recording. The origin (0, 0, 0) is at the small cuboid which reflects the placement of the transmitter.

3.1.3 Data recorded

In Figure 3.3 an example of the recorded data for one of the sensors is shown. The data is positional data of each limb recorded at a sampling frequency of 100Hz. This sampling frequency is more than sufficient to fully capture infant motion, as most similar studies use sampling frequencies of at least 50Hz (Marcroft et al., 2015).

The absolute positional data information is of little interest because this changes with each infant depending on their orientation and how far from the transmitter they are placed. Our interest is with movement patterns and behavior throughout the movement, and therefore we rely mainly on the velocity information which is obtained by numerically differentiating

and then smoothing of the positional data. This will be discussed in detail in section 3.4 of this thesis.

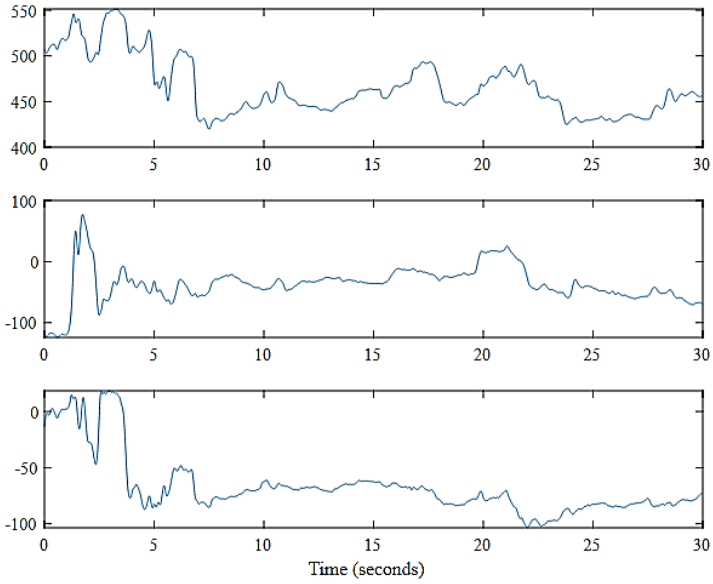

Figure 3.3 Example of data recording for one infant for one sensor (right arm). The subplots (top to bottom) are the x, y, and z axis respectively. In this plot, the x-axis is time (seconds) and the y-axis of each subplot is position (millimeters).

3.1.4 Infant age distribution

All of the infants recruited for this study were between the ages of 2 weeks old and 16 weeks old, as this is an age range where considerable development typically occurs during an infant's first year of life. Infants were typically called back for a second session that was within the second 8 weeks of life to again collect data that reflects any change in GMs. In total 49 infants were used in this study in which 27 infants attended for two recording sessions. In figure 3.4, the age distribution of each trial file is shown.

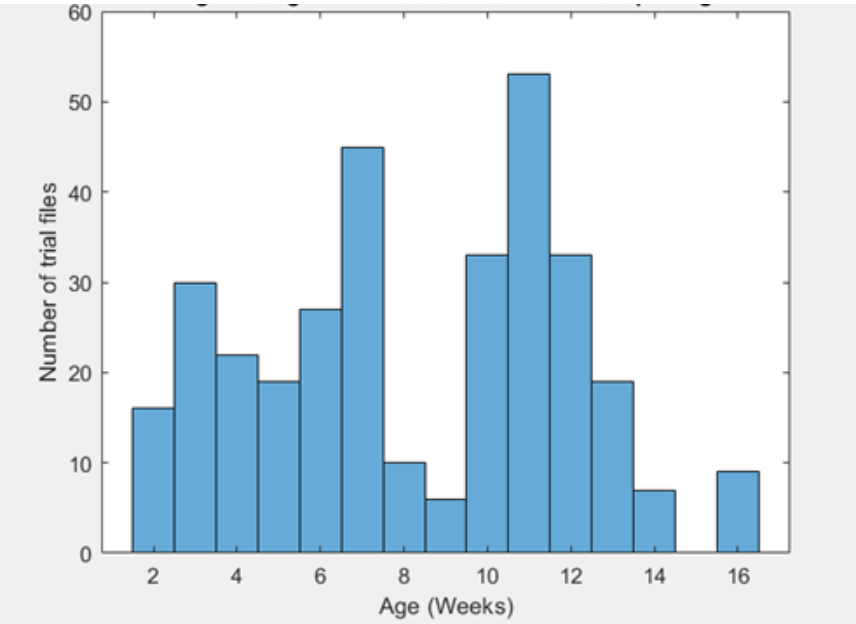

Figure 3.4 Age distribution of each trial file recorded. Each trial file was treated separately and was labelled with the infant's age in weeks

3.2 Data Pre-processing

3.2.1 Filtering

The positional data was filtered to contain only relevant frequency information. A zero-phase low-pass Butterworth filter was used with a cut-off frequency of 10Hz. Many studies apply a 5Hz-10Hz low-pass filter on movement data (Marcroft et al., 2015), so to avoid any attenuation in relevant frequencies, a 10Hz filter was applied. Each dimension of each sensor was filtered separately. The SNR of the signals was very high due to the accuracy of the Takstar tracking and so filtering was done mainly as a precautionary measure.

3.2.2 Data file segmentation

First, due to some movement artifact that typically shows up at the beginning or ending of a recording, the first few seconds and last few seconds of each recording are removed. Any recording that was less than 180 seconds was discarded from the analysis.

As most trials were between 3 – 5 minutes, we can split a recording into multiple recordings depending on time length and overlap used. This allows us to augment our dataset to increase the number of data points and to allow a more accurate training of classifiers later on.

One concern of splitting files into multiple files is that the training data no longer becomes exactly independent and identically distributed (i.i.d.), which is a central assumption in statistical learning theory (Hastie, Tibshirani, and Friedman 2009) and this negatively affects the generalization error of a classifier. However, as seen in figure 3.3, the GMs of an infant are typically non-stationary over a long recording. The behavior of the movements change and so we make the assumption with our analysis that over a trial, the movement features are sufficiently non-stationary so as to allow us to assume that even if a single file was split into multiple files, we still maintain approximately i.i.d. data.

Through trial and error, it was found that files 180 (seconds) in length with 60 (seconds) overlap produced enough training data while capturing enough information about infant GMs. For example, a 5 minute recording would be split into two files from [0 seconds to 180 seconds] and [120 seconds to 300 seconds]. This segmentation scheme gives us a total of 325 files to extract features from.

3.3 Feature Extraction

The features extracted from the GMs were carefully chosen to reflect aspects of complexity, fluency, and variability of GMs.

In our setting, one would have to be careful not to use features that might be biased by the physical size of the infant. Because we are predicting and clustering on age group it is natural that the size of the infant would bias certain features, and this would introduce size as a confounding variable into the model. For example, if we take the absolute range of motion of the infant as a feature, it will likely be larger for older infants compared to younger infants. Therefore we focus on features that are reflective of movement quality only.

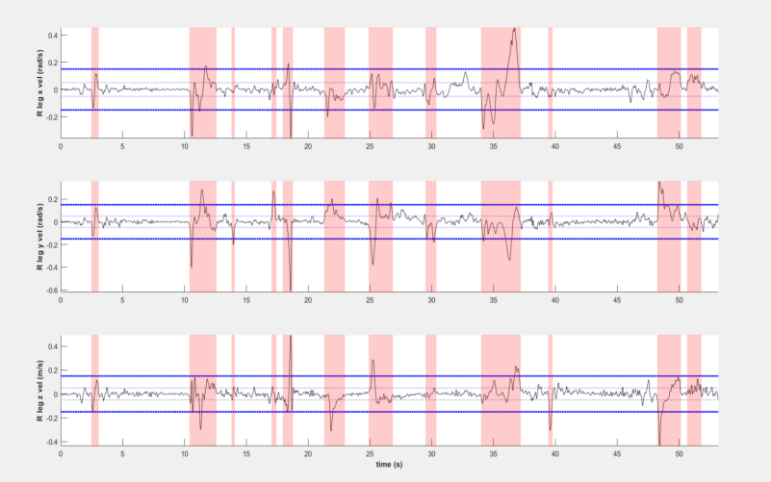

Figure 3.5 Movement 'units' or periods highlighted in pink. For each movement 'unit', some features are calculated and then averaged into a final feature vector. This plot is one sensor from one infant and the three subplots are the x, y, and z-axis respectively.

Some features were calculated as global features using the data from the entire recording. While other features were calculated and averaged over movement 'units' (Figure 3.5). These are defined as the period of time where the tangential velocity of the sensor exceeds a certain threshold and then falls back below it, defining a clear movement 'period' for each sensor.

3.3.1 Features modelling complexity

Complexity of movements are modelled by continuous values that are larger if there are more interactions between limbs during movement, or movements with relatively higher spatial exploration. To model these two aspects we used tortuosity, and absolute deviation of azimuth and elevation angles of the limbs.

Tortuosity (Eq 3.1) is defined as the ratio between the straight line length and the total curve length between the start and stop positions of a movement unit.

$$tortuosity = \frac{(x_f,y_f,z_f)-(x_s,y_s,z_s)}{arclen(movement)}, \quad (3.1)$$

Where *arclen(_)* takes the arc length from the start coordinates until the stop coordinates of the movement unit. This quantity has a maximum value of 1 if the limb moves in a straight line and then stops, and it has a minimum value close to 0 if the movement pattern swirls and curves before stopping. In other words, simple movements have a tortuosity close to 1 and spatially complex movements have a tortuosity close to 0.

Mean Absolute Deviation (MAD) is also used to quantify complexity by taking the MAD of the movement units' azimuth angle and elevation angle. The coordinates of the movements (x, y, z) are converted into spherical coordinates from which we extract the azimuth angle (parallel to the torso of the infant) and the elevation angle (perpendicular to the torso of the infant).

Finally the MAD of the distance between each combination of the four limbs (e.g. distance from right arm to left arm) is calculated over the entire trial. This quantity will have a higher value for infants who moved their limbs relative to one another in different directions, i.e. a limb's movement patterns were distinct to the other limbs.

3.3.2 Features modelling fluency

Fluency of movements are movements that have smooth acceleration and deceleration. This quantity is a little challenging to model due to competing definitions, but one of the best mathematical models was found to be the spectral arc length (SAL) of the movement units (Balasubramanian et al., 2012). For each sensor (table 3.1), we take the tangential velocity or speed $v(t)$ during a movement unit, and then the spectral arc length is defined as follows:

$$SAL = -\int_0^{\omega_c} \left[\left(\frac{1}{\omega_c}\right)^2 + \left(\frac{d\tilde{V}(\omega)}{d\omega}\right)^2\right]^{\frac{1}{2}} d\omega; \quad \tilde{V}(\omega) = \frac{V(\omega)}{V(0)}, \tag{3.2}$$

Here $V(\omega)$ is the Fourier magnitude spectrum of $v(t)$, ω_c is the cut-off frequency which we are measuring the arc length until. The $\frac{1}{\omega_c}$ term in the integral is a normalizing constant that normalizes the SAL with respect to the cut-off frequency and the magnitude as a whole

is further normalized with respect to the DC value at 0Hz. This value is averaged over all movement units into the final feature vector for the infant. Based on equation 3.2, the SAL of fluent movements will be relatively higher than low-fluency movements, because as demonstrated in (Balasubramanian et al., 2012), low-fluency movements have many more frequency components due to the uneven speed and quick direction changes which causes an increase in the arc length of the Fourier magnitude spectrum of the speed profile.

One other feature for fluency is how much the speed changes in a movement unit. This was calculated as follows:

$$\boldsymbol{\Delta} speed = \frac{v_{max} - v_{min}}{duration}, \tag{3.3}$$

The idea being that if there was a big difference between the max and min speed in a short duration, this can imply jittery or unsmooth movements.

3.3.3 Features modelling variability

Variability features are those features which model infants who continuously produce new movement patterns. This is the opposite extreme of an infant who repeats the same motion over and over again. The first feature is the inter-limb correlation. Here we binarize movement profiles into 1's and 0's to represent movement or no movement. Then the hamming distance is computed, where a hamming distance of 0 shows complete disagreement between movements, and a hamming distance of 1 shows movement in complete synchronicity. This is computed for each pair of limbs, i.e. between right arm and left arm, right arm and right leg, and so on.

Finally, the sample entropy or SampEnt (Richman and Moorman 2000), is calculated for the distance over time between each limb and the torso. The sample entropy is a measure of predictability of a signal given previous values. The idea being that, if the infant's GMs are variable, the infant continuously produces new patterns, giving a larger SampEnt value. Whereas an infant who continuously repeats a limb's movements' will have a low entropy value for the respective limb.

3.3.4 Summary of features extracted

In summary, the following are the features extracted from each trial recording:

Calculated for each movement unit for each sensor and then averaged into 1 value per sensor	Spectral Arc Length
	Δ Speed
	Tortuosity
	Elevation/Azimuth MAD
Calculated over entire trial	Between-Limb distance MAD
	Between-Limb correlation
	Limb-Torso distance Sample Entropy
	% of time all limbs moving

Table 3.2 Features extracted from the trial files after pre-processing. In total there was 37 features extracted.

3.4 Complete Pipeline

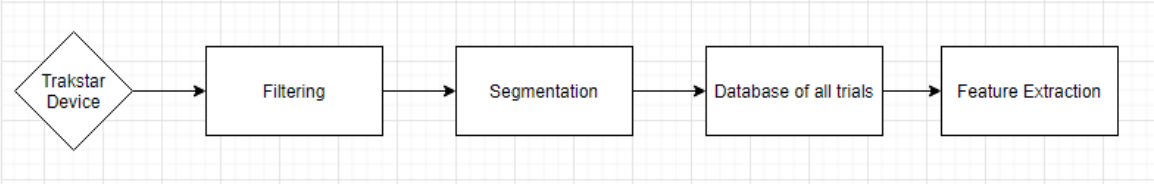

Figure 3.6 A simple figure illustrating the pipeline until the point of having a full feature table

It should be noted that having multiple trials per infant and segmenting long trials into shorter files to augment the dataset size, in our final feature matrix multiple rows can belong to the same infant. This potentially creates a correlation between rows in the feature matrix and we must be extra careful during validation to take this into account because it can overestimate the accuracy giving abnormally high results.

Once all the features are extracted, the data is ready to be passed into a classifier. In the next section we discuss the labels applied to the feature vectors and classification algorithms used to classify this data.

Chapter 4

Supervised Learning

Our main goal in this section is to train a classifier that can accurately classify infants into their age group given their movement features. At this point we have our feature matrix ready, which has the dimensions [325 x 37] and where each row represents a data point each of which is characterized by 37 features as described in table 3.2. In the machine learning literature, there are many different options for classifiers, ranging from simple linear separators to highly non-linear functions like deep neural networks (Hastie, Tibshirani, and Friedman 2009). To narrow down the choices of classifiers, we look at the constraints of our problem such as the dataset size, the feature dimensionality, and interpretability for the application of studying infant kinematics. In this thesis, our requirements of a classifier are as follows:

- Can be trained on a relatively small dataset without overfitting
- Interpretable by giving feedback on which features had the highest impact on the classification decision
- Making no assumptions on the distribution of the data in feature space

To explain the first point, one would have to be careful with smaller datasets not to use classifiers that have a large number of degrees of freedom to fit the data. This is because a smaller dataset is not necessarily a good representation of the general population that we are trying to generalize to, and so any classifier that fits strongly to a smaller dataset will have a lower probability of generalizing well to the population. This is known as the bias-variance trade off (Hastie, Tibshirani, and Friedman 2009). For example, an artificial neural network or a decision tree has very high variance but low bias because they don't bias a classifier towards a specific type of function and so the decision boundary can look very different across different samples of the data. Whereas an SVM with a linear kernel (Hsu et al., 2003) would have a relatively higher bias but lower variance because the decision boundary will not change significantly if it is trained on different samples of the data. In our case due to having a smaller dataset, classifiers with higher inductive bias (such as a linear classifier) are safer to use to avoid overfitting to the data. One other type of classifier is used to better explore the data, which is an ensemble decision tree classifier known as the random forest (RF) classifier (Breiman., 2001). Though the decision boundary of such a classifier is very non-linear (compared to an SVM), typically ensemble methods average their classification result across many "weak" classifiers which has the effect of reducing prediction variance and thus reducing overfitting (Breiman., 2001). To address the second requirement above, both the linear SVM and the RF have the added benefit of giving feature weights or feature importance, giving us insights into which features of GMs have the strongest discriminative power to classify age group.

4.1 Labels

The age of each infant (in weeks) was recorded at the time of recording the infant's GMs. As seen in figure 3.4 the ages range between 2 weeks old to 16 weeks old. Due to high variability in the infant's development, regressing the age on the movement features will likely not work, as not every 8 week old behaves as other 8 week olds. A more logical approach would be to bin the ages into a different number of groups and classify the group as a whole.

4.1.1 Binning for two age groups

The first binning is into two age groups, as this naturally reflects the taxonomy of normal GMs (table 1.1). The infants, or more accurately the trial files recorded, were split into two groups, the files of infants equal to or less than 8 weeks in age, and those above 8 weeks old in age.

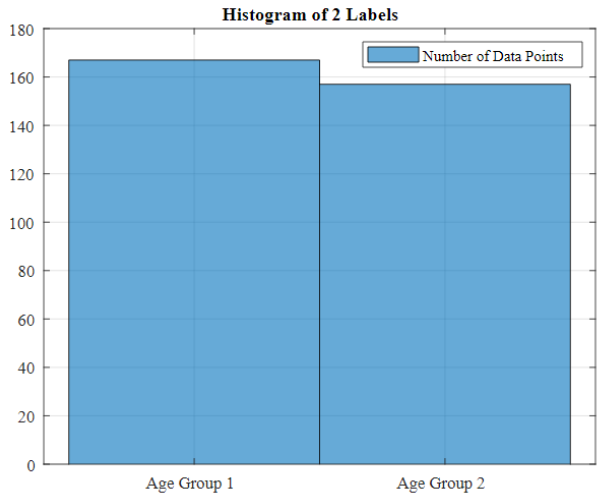

Figure 4.1 Histogram of labels for two bins. Table 4.1 shows the age range break down.

Age	Label
0 – 8 Weeks Old	1
9 – 16 Weeks Old	2

Table 4.1 Age range and corresponding label assigned

4.1.2 Binning for three age groups

A finer binning of the ages is also conducted that would reflect the transitionary age group between the ages of 6 weeks and 10 weeks old. A more detailed binning like this allows greater insight to be derived from the data, as having the ability to further classify a third group within the data can show a finger developmental trajectory for the healthy infants.

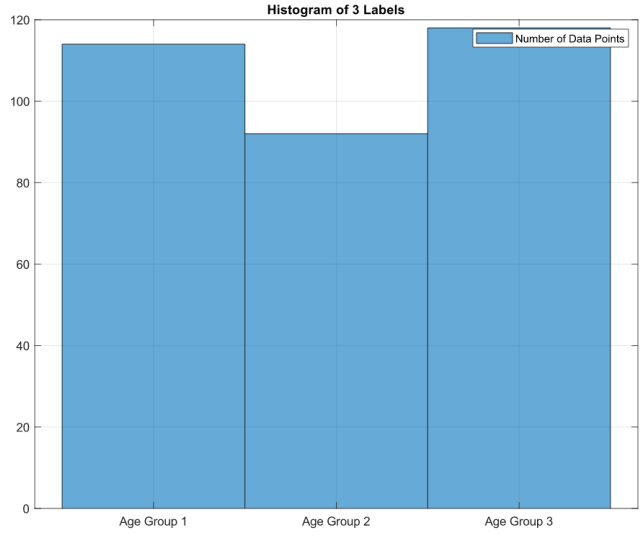

Figure 4.2 Histogram of labels for three bins. Table 4.2 shows the age range break down.

Age	Label
2 – 6 Weeks Old	1
7 – 10 Weeks Old	2
11 – 16 Weeks Old	3

Table 4.2 Age range and corresponding label assigned

4.2 Cross-validation

Before training the classifiers on our feature matrix, it is crucial that a correct validation scheme is selected. Validation in this context refers to evaluating the generalization error of the classifier that is trained on our data. In our feature matrix, multiple rows may belong to the same infant (section 3.2.2) but we are interested in estimating the generalization error to new infants. Therefore we must be careful to be evaluating the classifier on data points that would represent completely new data points not seen by the classifier during training.

4.2.1 Types of cross-validation

There are two main types of validation schemes known as the train-test split and the 'k'-fold cross-validation (Hastie, Tibshirani, and Friedman 2009) . For the train-test split, we randomly select 75% of the rows in our feature matrix, train the classifier on these selected rows, and then test the accuracy on the remaining 25% of the rows. In the k-fold cross validation scheme, a loop is setup inside which we randomly choose some data points to be left out (e.g. 5% of rows), a classifier is trained and then tested on the left out data points. This is repeated 'k' times and the test error is averaged to give a final estimate. However

for our setting neither validation scheme will exactly work because multiple rows of the feature matrix can belong to one infant, and this will show a falsely large generalization accuracy because we are training and testing the same infant on themselves. Another issue is that some infants actually behave as infants from a different age group. If these infants were in the testing set this can give the false result of a low generalization accuracy when the source of error was the infant being an outlier in their group.

4.2.2 Leave-k-subjects-out validation

To solve the problem of multiple rows belonging to the same infant, we adopt an n-fold cross-validation scheme, but instead of randomly leaving out rows, instead on each loop iteration we choose k infants, leave out all the rows belonging to these k infants, and then test on them. This is repeated 'n' times to train and test on as many different combination of infants as possible (Figure 4.3).

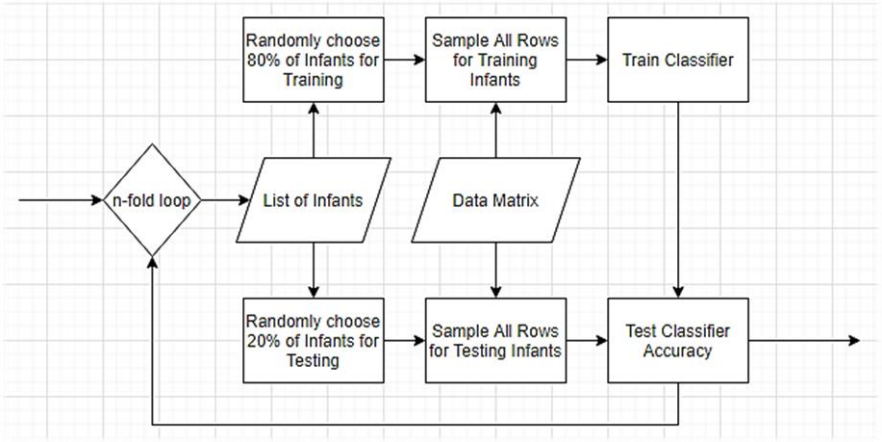

Figure 4.3 Leave-k-Subjects Out example flowchart where k is 20% of all infants repeated n times

4.2.3 Adjusted Accuracy and Confusion Matrix

For the LkSO validation loop, two main statistics will be reported in the results section. The first is the adjusted accuracy statistic which takes into account class imbalance while calculating the accuracy. The second is a confusion matrix percentage estimate.

The adjusted accuracy is calculated as follows (Mosley., 2013):

$$acc_{adj} = 0.5\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right), \tag{4.1}$$

Where TP is true positive, TN is true negative, FN is false negative and FP is false positive. Because the left out infants are randomly sampled, we cannot guarantee that the test set has balanced classes on each iteration of the validation loop, and therefore accuracy may become inflated due to class imbalance. The adjusted accuracy addresses this by weighing each sample by its class's prevalence, giving a true reflection of whether the classifier is either randomly guessing or always predicting the prevalent class, or whether it truly learned to distinguish between the labels. In the multiclass case the adjusted accuracy is calculated in a one-vs-all fashion for each class and then averaged over all classes.

A confusion matrix is a table that compares predictions with the true class labels for each class. The confusion matrices were calculated during the LkSO loop by calculating the percentage of misclassified files on each LkSO iteration and then finally normalizing all the confusion matrix rows to 100%.

4.3 Linear Classifier

The first classifier is a linear SVM which creates a linear boundary between each class. Given a training dataset (as per 4.2.2), we use the *sklearn.svm.LinearSVC* class from the Sci-kit Learn library implemented in Python (Pedregosa et al., 2011). The classifier is trained on both the 2 bin labels (4.1.1) and 3 bin labels (4.1.2) and the results of both analyses are reported below.

4.3.1 Data preparation

Before the data matrix can be passed onto the SVM classifier, a scaling step must be conducted in order to normalize every feature. This is important because features which have a larger scale will dominate the smaller scaled features thus removing their discrimination power although it may be the case that the smaller scale feature itself is a better feature (Hsu et al., 2003). Each feature column in the feature matrix is scaled according to:

$$\tilde{x_i} = \frac{x_i - mean(x_i)}{std(x_i)}, \tag{4.2}$$

4.3.2 SVM hyper-parameters

The support vector machine is an algorithm that classifies a data point according to which side of the decision boundary the point falls on. Ideally there exists a natural separation in the feature space between the classes but in practice that is rarely the case. For an SVM

with a linear kernel, the hyperparameter that controls the level of overlap is 'C' (Hsu et al., 2003).

As for the kernel, we used a linear kernel to avoid increasing the degrees of freedom of the model. Having too many hyperparameters to tune can easily result in overfitting and our validation loop is built for estimating generalization error and not estimating hyperparameters. A few values were handpicked and tried for the SVM hyperparameter 'C', which controls the strength of regularization and allows a certain number of misclassifications in exchange for better generalization error. The end results did not change significantly for the handpicked value of C = 5 and so this value was used. For a larger dataset, a grid search loop for the hyperparameter would first be used and then the generalization error would be estimated separately. For the 3 class case the linear SVM was trained in a one-vs-all fashion in which 3 hyperplanes are constructed for each class, and for the respective class it was labelled 1 and all other classes -1 and then the classifier was trained. This creates 3 hyperplanes one for each class.

4.3.3 SVM classification results

The results of the LkSO validation will be reported for the 2 class binning (table 4.1) and then for the 3 class binning (table 4.2).

The accuracy on the left out infant files was calculated over 300 iterations and the results for the 2 bins case are shown in the figures below. Note that all accuracies are reported as adjusted accuracies. A 0% adjusted accuracy meant that the classifier accuracy on the age group of the left out infants was no better than random chance (i.e. nothing interesting was

learned), and a 100% adjusted accuracy meant that the classifier correctly classified all left out infants. The confusion matrices were normalized to have rows adding to 100% indicating percentage of correctly and incorrectly classified rows.

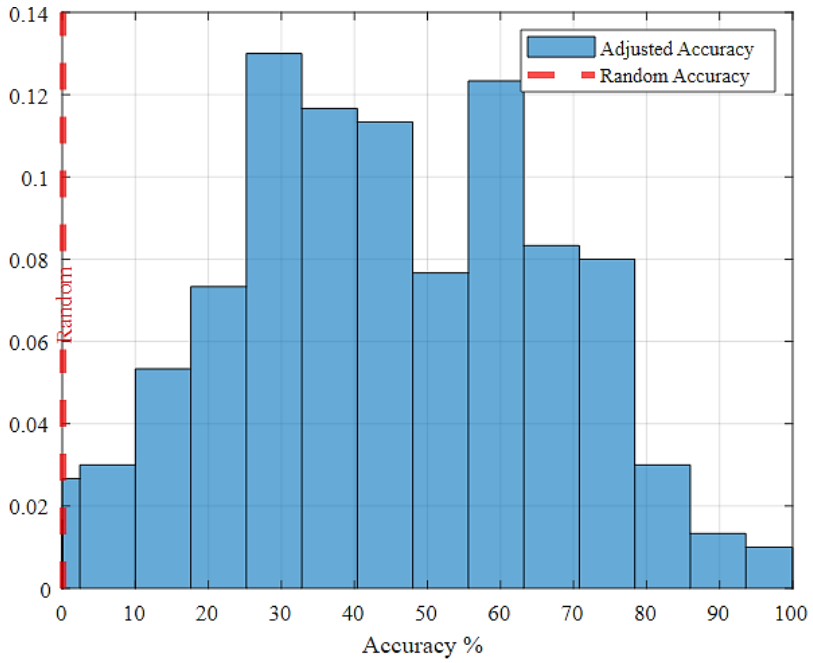

Figure 4.4 SVM adjusted accuracy scores on LkSO iterated 300 times for 2 classes. The average value being around 50% adjusted accuracy which is well-above a randomly guessing classifier. The y-axis is the relative probability of an accuracy percentage.

Next the confusion matrix for the 2 class SVM classifier is shown (figure 4.5). The calculation of the confusion matrix in the LsKO is done according to section 4.2.3 and so each row of the confusion matrix is normalized to 100%.

The confusion matrix (figure 4.5) indicates that the classifier was able to correctly classify about 70% of the left out infants which is a high accuracy value given that very little hyperparameter tuning was conducted and the model is a linear separator.

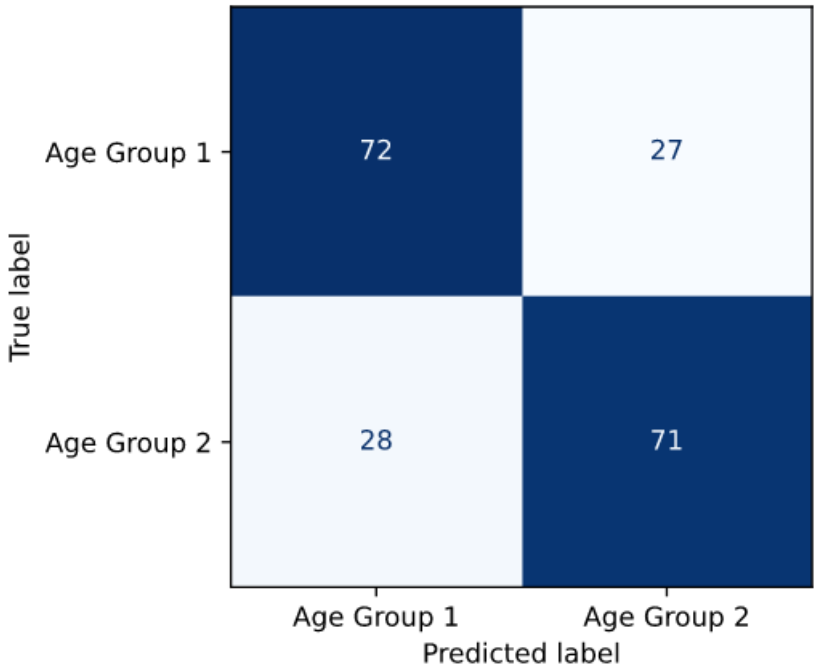

Figure 4.5 SVM confusion matrix of the 2 class classification tested as an LkSO validation scheme. The entries are percentages where each row adds to approximately 100%.

The adjusted accuracy (figure 4.4) shows an approximately bimodal distribution indicating that for some infants, the classifier was less certain about their correct classification while for others it had a higher degree of certainty.

Similarly for the 3 class binning (section 4.1.2) the adjusted accuracy and confusion matrix is reported. Compared with the adjusted accuracy of the 2 class binning, the adjusted

accuracy of the 3 class binning had a lower average value and was unimodal with values closer to 0% (Figure 4.6).

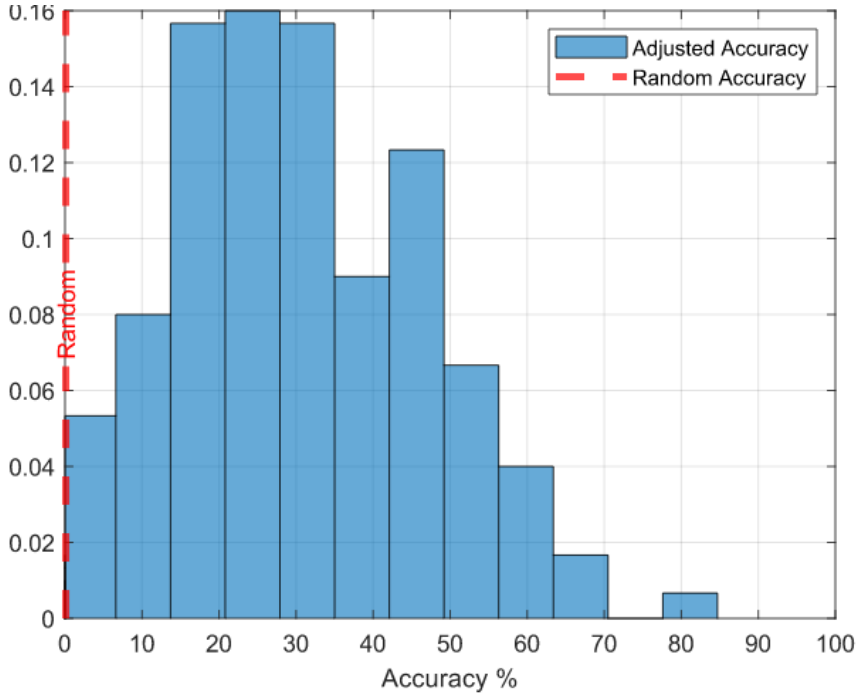

Figure 4.6 SVM adjusted accuracy scores on LkSO iterated 300 times for 3 classes. The average value being around 40% adjusted accuracy which is well-above a randomly guessing classifier but less than the 2 class case. The y-axis is the relative probability of an accuracy percentage.

This shows that for a finer binning of the age groups, the precision becomes negatively affected. A better visualization of the misclassifications can be seen in the confusion matrix (Figure 4.7). Here it is clear that between adjacent age groups physiological behavior overlaps and especially for age group 2 (between the ages of 6 weeks and 10 weeks) the infants tend to be misclassified on both sides. This behavior of misclassification is expected

because as noted in (Einspieler and Prechtl 2005) age group 2 marks a transitionary change between one type of GMs into another type (see table 1.1).

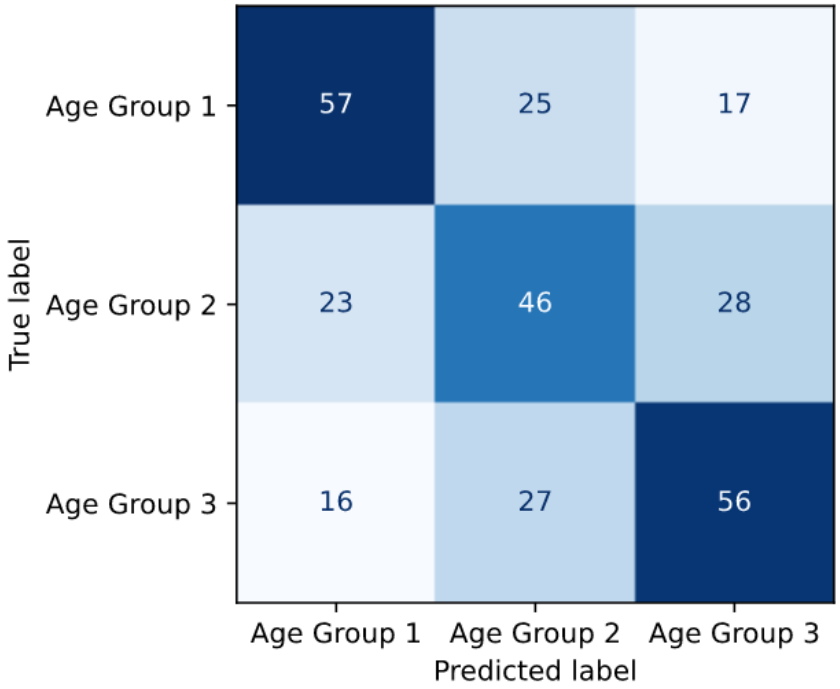

Figure 4.7 SVM confusion matrix of the 3 class classification tested as an LkSO validation scheme. The entries are percentages where each row adds to approximately 100%. The results indicate expected overlap between adjacent age groups and therefore reduced precision.

4.3.4 Feature importance

For a linear kernel SVM, we can directly derive the weights of the features from the SVM itself. This is because for a linear kernel, the separating hyperplane solution is defined in terms of weights of features of the inputs. And since all features are rescaled before training the SVM, the weights can be interpreted as the importance of each feature. In figure 4.8 the top 10 feature weights for the 2 class SVM are selected. Note that we are only interested in absolute values and so we only look at the magnitude of the feature weights.

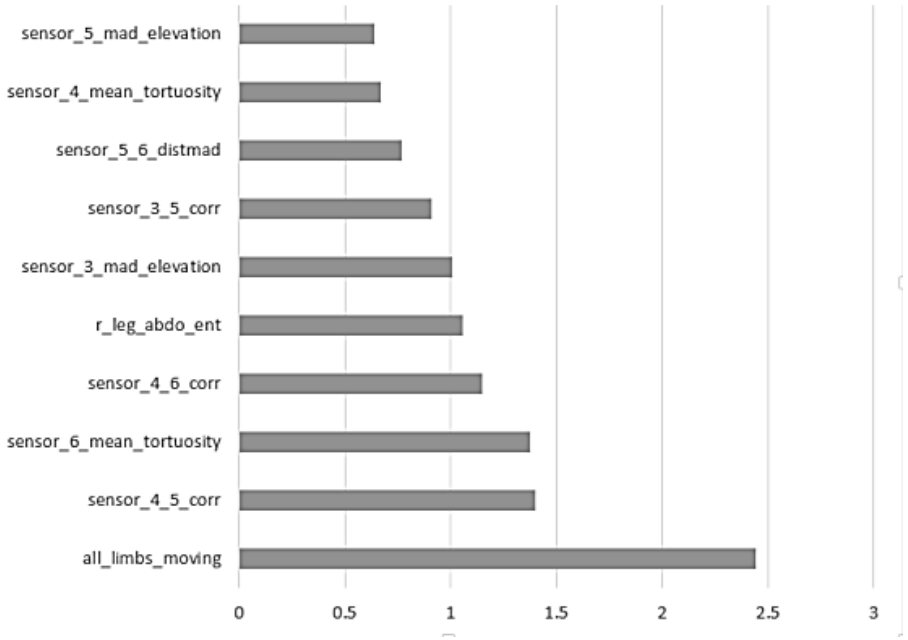

Figure 4.8 Top 10 average feature weights for the 2 class trained SVM

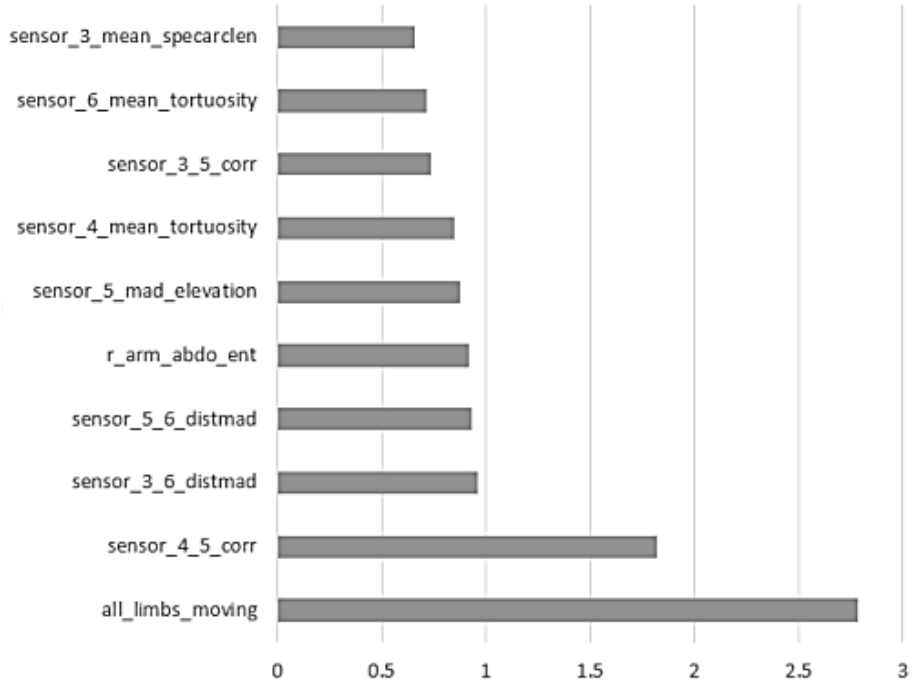

Figure 4.9 Top 10 average feature weights for the 3 class trained SVM

All_limbs_moving (1 feature)	% of trial where the infant is moving all their limbs at the same time
Sensor_3_5_corr, Sensor_4_5_corr, Sensor_4_6_corr (3 features)	Timing correlation between: Right Arm and Right Leg, Left Arm and Right Leg, Left Arm and Left Leg
Sensor_3_6_distmad, Sensor_5_6_distmad (2 features)	Mean deviation of distance between: Right Arm and Left leg, Right Leg and Left Leg
R_arm_abdo_ent, R_leg_abdo_ent (2 features)	Sample entropy of the distance between: Right Arm and Abdomen, Right Leg and Abdomen
Sensor_3_mad_elevation, Sensor_5_mad_elevation (2 features)	Mean deviation of elevation angle of: Right Arm, Right Leg
Sensor_4_mean_tortuosity, Sensor_6_mean_tortuosity (2 features)	Tortuosity of: Left Arm, Left Leg
Sensor_3_mean_specarclen (1 feature)	Spectral Arc Length of: Right Arm

Table 4.3 List of the top 10 features combined from both the 2 class and 3 class SVM, and organized into categories of the features. Features names (left) and descriptions (right).

4.4 Random Forests

Another classifier that was trained was the random forest (RF). This classifier is well-known and is universally used in many different applications (Zhang and Ma 2012). While a single decision tree is easy to interpret because we can trace the decision process, it can easily over fit especially to a smaller noisy dataset. To address this limitation, we look at random forests. A random forest is essentially an ensemble of many decision trees that trains each tree on a randomly chosen subset of observations and features (Breiman., 2001). This ensembling scheme creates many "weak" learners because each tree is being trained only on a randomly chosen subset of the data for both rows and columns. However the final classification result is the averaged or aggregated result from all the weak ensembles, the idea being that many weak learners create stable accurate predictions. Averaging results from weak classifiers is also a standard technique for reducing overfitting to data (Hastie, Tibshirani, and Friedman 2009).

One major benefit of RF is that it gives a rather straightforward framework to calculate feature importance based on information gained within each decision tree. In section 4.4.3 we give a brief overview of how this is done.

4.4.1 RF parameters

For the RF implementation, we use the *sklearn.ensemble.RandomForestClassifier* class from the Sci-kit Learn library implemented in Python (Pedregosa et al., 2011). The

classifier is trained on both the 2 bin labels (4.1.1) and 3 bin labels (4.1.2), and the results of both analyses are reported in 4.4.2.

The main parameter of RFs is the number of decision trees in the forest. As with the case in the linear classifier section, because of how our LsKO validation loop (section 4.2.2) is setup, and due to the dataset size and variance, a separate hyper-parameter tuning procedure without overfitting to the data was not possible. Luckily, for RFs, the number of decision trees to use asymptotically does not lead to more overfitting and this is in essence due to the ensembling and randomness employed in training each decision tree. For a more detailed analysis of the asymptotic behavior of this parameter see Breiman (2001). A few randomly picked values were tried for this parameter, and the best results seemed to be achieved with around 100 decision trees.

In the Sci-kit Learn implementation we used, there are many other parameters for RFs, but these were all left at default values, because changing them did not significantly affect our final results.

4.4.2 RF classification results

The accuracy on the left out infant files was calculated over 300 iterations and the results for the 2 and 3 classes case are shown in the figures below, respectively. Note that all accuracies are reported as adjusted accuracies. A 0% adjusted accuracy meant that the classifier accuracy on the age group of the left out infants was no better than random chance (i.e. nothing interesting was learned), and a 100% adjusted accuracy meant that the classifier correctly classified all left out infants. The confusion matrices were normalized

to have rows adding to 100% indicating percentage of correctly and incorrectly classified rows.

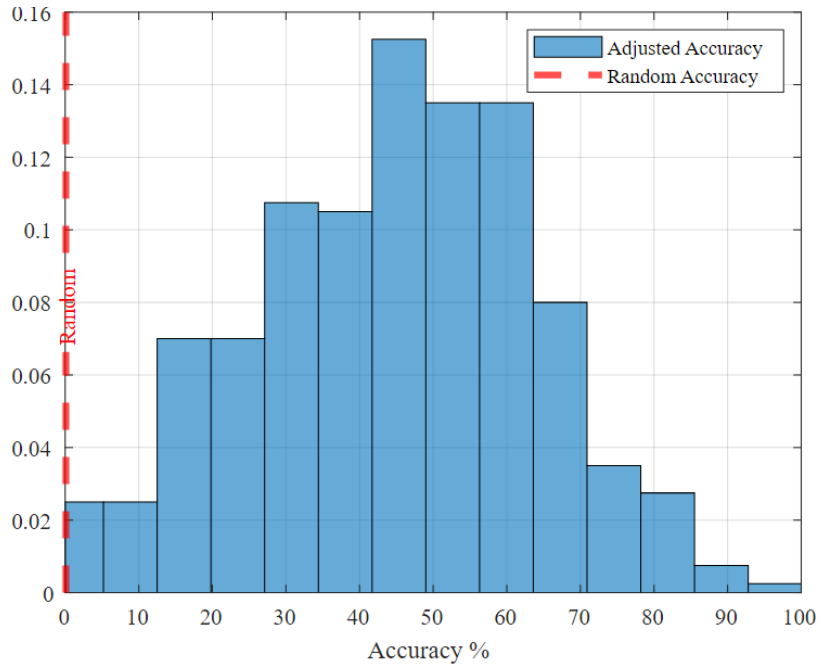

Figure 4.10 RF adjusted accuracy scores on LkSO iterated 300 times for 2 classes. The average value is approximately 50% adjusted accuracy, which is well-above a randomly guessing classifier. The y-axis is the relative probability of an accuracy percentage.

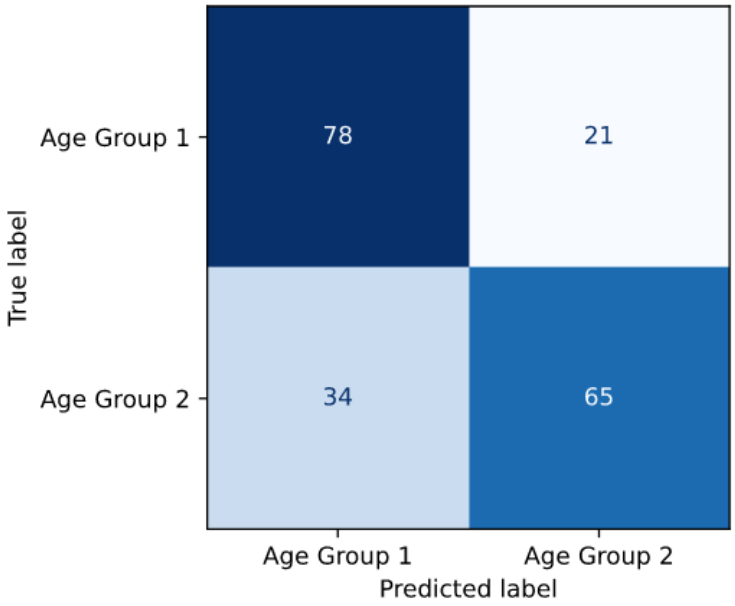

Figure 4.11 RF confusion matrix of the 2 class classification case tested as an LkSO validation scheme. The entries are percentages where each row adds to approximately 100%.

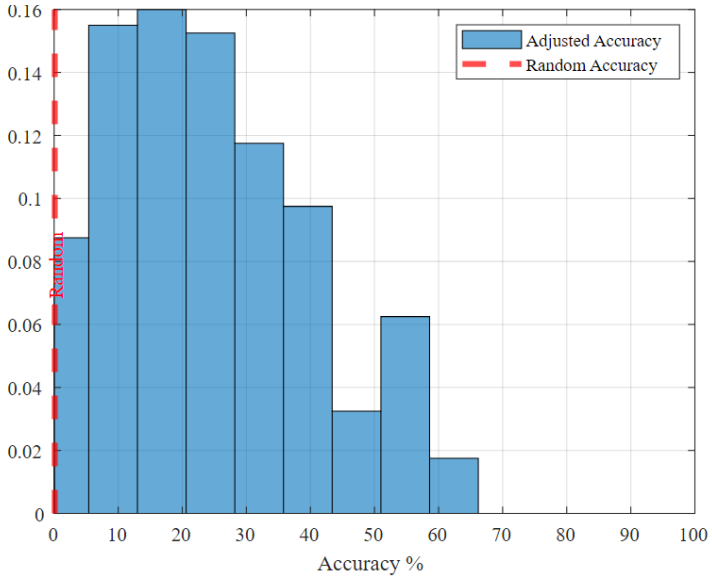

Figure 4.12 RF adjusted accuracy scores on LkSO iterated 300 times for 3 classes. The average value of approximately 40% adjusted accuracy is well-above a randomly guessing classifier but less than the 2 class case. The y-axis is the relative probability of an accuracy percentage.

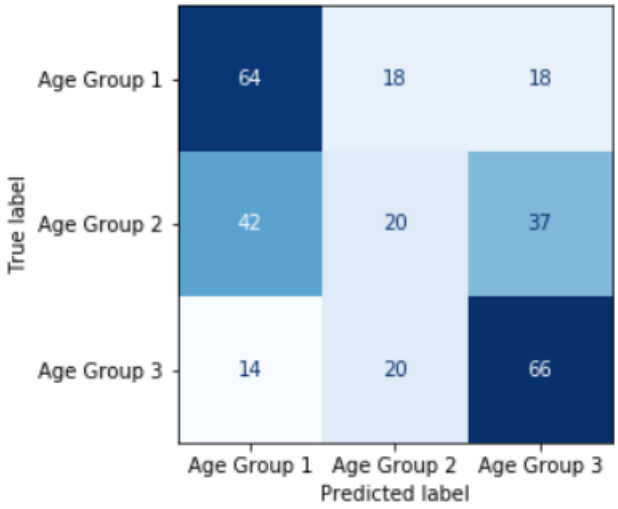

Figure 4.13 RF confusion matrix of the 3 class classification tested as an LkSO validation scheme. The entries are percentages where each row adds to approximately 100%.

4.4.3 Feature Importance

A RF is a large collection of decision trees, where each decision tree, independent of one another, is trained on randomly chosen rows and columns from the original feature matrix. Therefore for each decision tree, there are observations which are not used to train the decision tree (also commonly referred to out-of-bag (OOB) observations), and the OOBs can be used to estimate generalization error for that specific tree. Using this idea, feature importance can then be computed by random shuffling all the values in a feature column, and then checking how that affects the accuracy of the relevant decision trees. This allows us to quantify the importance of each feature column across the entire RF. Due to the significant drop in accuracy for the 3 class RF (figure 4.13), the feature importance for that model will not be used for analysis, and therefore we only present the feature importance for the 2 class RF (figure 4.14).

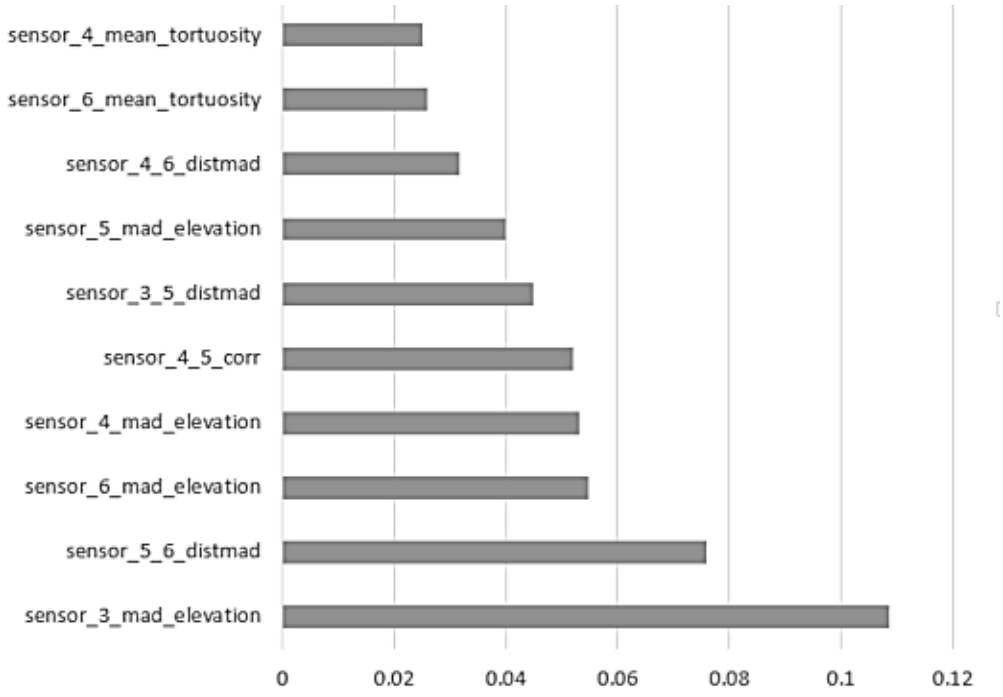

Figure 4.14 Top 10 average feature weights for the 2 class trained RF

Sensor_4_5_corr (1 feature)	Timing correlation between: Left Arm and Right Leg,
Sensor_4_mean_tortuosity, Sensor_6_mean_tortuosity (2 features)	Tortuosity of: Left Arm, Left Leg

Table 4.4 List of the top 10 features combined from both the 2 class and 3 class RF, and organized into categories of the features. Features names (left) and descriptions (right).

Table 4.4 (continued)

Sensor_3_mad_elevation, Sensor_4_mad_elevation, Sensor_5_mad_elevation, Sensor_6_mad_elevation, (4 features)	Mean deviation of elevation angle of: Right Arm, Left Arm, Right Leg, Left Leg
Sensor_3_5_distmad, Sensor_4_6_distmad, Sensor_5_6_distmad (3 features)	Mean deviation of distance between: Right Arm and Right Leg, Left Arm and Left Leg, Right Leg and Left Leg

4.5 Discussion of classification results

Comparing the SVM results with the RF results shows that there is a drop in precision and accuracy for the RF model, particularly for the 3 class classification task. This is somewhat expected a priori because the decision boundary of RFs are significantly more complicated (despite ensemble effects) compared to a linear SVM that just sets up a plane of separation between the data points. The RF therefore can be prone to fitting the noise in the data more strongly. However, there is another reason that most likely explains this drop in accuracy and this is to do with how a RF gains its strength as a classifier. A crucial aspect of training the decision trees in a RF is that the trees must be independent of one another (Breiman., 2001). However, because we have multiple rows belonging to the same infant in our feature

matrix, this can create a correlation between the rows. The decision trees in turn are separately trained on these correlated rows, which can reduce the independence between the decision trees and therefore negatively affect the generalization capability of the RF. An analysis of subject-level bootstrapping for RFs is conducted in Karpievitch et al. (2009), and it was shown that a RF's accuracy can be improved by ensuring that the decision trees are trained on random rows that belong to strictly different subjects, but this is left for future work to test on our dataset.

Of most importance to these results is that the simpler model (SVM) was capable of separating the classes quite accurately. Moreover for the 3 class case, class 2 shows significant overlap into class 1 and 3. As these are consecutive age groups this overlap is expected. In fact, class 2 are the infants in the age group 6 – 10 weeks old and Einspieler and Prechtl (2005) speaks about this as the transitionary group that will show aspects of both groups as the infant begins changing their GM patterns. This shows that our classifier is picking up on this transitionary period and we can conclude that the classifier is learning something clinically interesting.

It must also be mentioned that there are infants who were group 1 but were misclassified into group 3 and vice-versa. There is quite a large gap in age between these two groups and an in-depth investigation into the infants must be conducted, though this is left for future work. Some speculation can be done, and there are many reasons why this could happen, but most likely it is due to infants behaving abnormally due to fussiness or at the other end of the spectrum, sleepiness. It could also be that a certain movement feature value caused

them to become misplaced and so an investigation into interpretable decision models can bring to light the reason for misclassification.

4.6 Discussion of feature importance

More interesting than the classification results are the selected features. Using two different algorithms, we derived feature importance or feature weights, which were most important for the classifier's decision. The top 10 features were selected from the 2 and 3 class SVM and only the 2 class RF model. There is a lot of overlap in the selected features, but more interestingly, most of the features belonged to the complexity or variability modelling features but only 1 fluency modeling feature was selected (see section 3.3 for definitions). The fluency feature selected was the SAL of the right arm and this makes sense because at the ages of infants which were recorded, the arms perform more movement relative to the legs and so the infant has a higher chance of displaying fluent movements in the arms not the legs. It is known from the GM literature that fluency begins developing towards the latter part of the second 8 weeks of life and therefore more likely to in the very oldest infants in our sample population. In fact those approaching their fifth month (Einspieler and Prechtl 2005). And so it is conceivable that fluency might not be important to discriminate the age groups in our age range and in fact that would agree closer with the GM literature. For the SVM classifier, the feature *'all_limbs_moving'* had the highest weight for both the 2 and 3 class tasks. This feature is associated with complex movement, if all limbs are simultaneously being used then that would be qualitatively more complex

than moving a single limb. As for the RF classifier, its most selected feature category was the MAD of the elevation angle of each of the four limbs. This shows that the amount of vertical movements (anti-gravity) is a sign of healthy development.

Conclusion

5.1 Future Work

Below we outline some of the immediate next steps that can be taken with this project. These include both clinical exploration but also algorithmic exploration, using different algorithms to derive more insight from the data.

5.1.1 Clustering and Dimensionality Reduction

As per the discussion in Chapter 4, it was shown that our dataset exhibits some variance, and labels based on a finer binning of age were difficult to accurately predict with adjacent age groups showing overlap (figure 4.7 and figure 4.13). Therefore a natural next step is to apply clustering techniques that allow the data points to fall into their natural clusters without imposing strong labels on each data point. We hypothesize that the clustering should demonstrate some separation based on age, but observations will overlap between clusters due to some infants developing faster/slower than others in their age group.

There are many clustering techniques that exist in the literature with various assumptions on feature distributions, outliers, dimensionality, and so on. Initially the Gaussian Mixture

Model (GMM) was chosen due to its ability to model various shaped clusters. And since we do not assume independence of our feature columns (in fact several are correlated), the GMM is an ideal clustering solution to use.

The Gaussian mixture model is a ubiquitous clustering algorithm that can model a wide range of data distributions. The model is given by the following equation:

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k), \quad (5.1)$$

Where K is the number of Gaussian clusters, π_k is the proportion out of 1 of each Gaussian cluster, μ_k are the means of each Gaussian cluster, Σ_k are the covariances of each cluster, and N is the normal distribution. The parameters of this model are fit to a data set using the Expectation Maximization (EM) algorithm (Bishop., 2006). Note that because the dimensionality of our data is [325 x 37] and multiple rows are sampled from the same infant, this can reduce the accuracy by which the GMM parameters are estimated. Because we want to impose as little structure on the parameters as possible, this problem requires that we estimate a covariance matrix for each cluster that has d^2 parameters where d is the number of features in our problem. This causes the number of parameters of the model to estimate to grow very quickly. Reducing the number of parameters to be estimated from a dataset can lead to more accurate parameter estimations, and this leads us to the Mixture of Factor Analyzers (MFA) models (McNicholas and Murphy 2008).

Mixture of Factor Analyzers are based on factor analysis models, and are combined with GMMs. Factor analysis is a form of dimensionality reduction that models correlations that exist in the variables of the data by modelling the features as linear combinations of hidden

factors. For MFA, the covariance matrices Σ_k are parametrized by lower dimensional matrices with dimensions given by [p x q], where q << p (Ghahramani and Hinton 1996). This effectively reduces the number of parameters that need to be estimated from the data. We refer the reader to Ghahramani and Hinton (1996) and McLachlan and Peel (2000) for an in-depth discussion of these models. The point of interest for our application is the reduced number of parameters compared to GMMs, and thus giving better estimates for the clustering parameters.

For this task we use the R package *pgmm* (McNicholas et al., 2019). This package trains 12 families of MFAs with various constraints imposed on the parameters as shown in McNicholas and Murphy (2010). The best model is selected based on two criteria: Its reflection of age groups and the relative Bayesian Information Criterion (BIC) of the model. The BIC balances model fit with the number of parameters used. This is because a trivial solution is to fit a separate cluster for each data point but of course this does not tell us anything interesting about the data. The BIC has been shown to work well for model selection in the case of GMM and MFA (Steele et al., 2010).

One major issue with EM algorithms is the parameter initialization because part of the algorithm is to initialize the parameters and then to iteratively update them according to data fit. Depending on how this is done, the solution can converge to clustering solutions that are wildly different from one another (Bishop., 2006). In our application, interpretable solutions are crucial to allow applicability in a clinical setting. Running on a completely unsupervised case with no guidance from age labels leads to clusters that might not reflect

any interesting patterns, particularly because not all features in our feature matrix provide useful information for grouping based on age.

In the *pgmm* R package (McNicholas et al., 2019), their algorithm supports passing in group membership information for some of the data points. This acts as a sort of prior on the cluster assignments for the data points, and this helps guide the optimization to converge on clusters that reflect the different age groups. As mentioned previously, not every infant truly behaves according to the label assigned to them, some infants behave as older infants do and vice-versa (figure 4.7)

With this in mind, we adjust our code to pass in some cluster assignments with our data points. All respective rows of 30% of infants in each age group are kept with their originally assigned label based on their age in weeks, while the remaining rows of the other 70% of infants in each age group is given a label of '0' meaning that their group membership is unknown and clustering should assign them. We still expect most of the unlabeled infants in each age group to be assigned into their respective age group, but the clustering should do this with no label information.

The rest of the inputs into the *pgmm* algorithm are the range of number of clusters to use, and a range of the number of factors to use. In our case the range of number of clusters was set to [3 to 8] clusters and the number of factors 'q' was [3 to 8] as well.

The results of the clustering algorithm are shown in figure 5.1. The *pgmm* clustering algorithm, based on the BIC criterion, chose the 'CUU' model with 3 clusters and 8 factors, significantly reducing the dimensionality of our data. This model constrains one of the

parameters, referred to as the loading matrix, to be shared across clusters. For further details and mathematical background see McNicholas and Murphy (2008).

Figure 5.1 Cluster assignment compared with original age group label (based on table 4.2)

We also report for the infants who are in age group 2 that were assigned to clusters 1 and 3, their median ages are shown in table 5.1. This shows that age group 2 truly is a transitionary period where the younger infants in age group 2 tend to cluster with age group 1 and the older infants in age group 2 cluster with age group 3.

Cluster 1	7 weeks old
Cluster 3	10 weeks old

Table 5.1 Median age of age group 2 infants were clustered with the other age groups instead

With this clustering scheme we see that despite not knowing the true age group of 70% of the infants, the clustering algorithm placed many of these infants in the cluster that contains other infants who, in the majority, are of the same age group as one another.

Of particular interest are infants who are in age group 3 that were clustered with age group 1. Referring back to the notes during the recording sessions we see comments from the clinician that say that these infants were “fussy” and had “insufficient complexity & variability”. Indicating that relative to the older age group, the younger age groups had less complexity and variability in their movements.

The infants who were in age group 1 and 2 who clustered in cluster 3 are those infants who were relatively more active and who utilized all their limbs in a complex fashion. Particularly one infant who received the highest GM rating from the clinician, clustered with age group 3 despite their age falling into age group 2.

Again it must be emphasized that, though these results do validate our expected clustering, they cannot be completely trusted because of how the data was sampled. Multiple observations were sampled from the same infant and this breaks the i.i.d. assumption of statistical models such as the GMM and by extension the MFA. More sophisticated clustering algorithms that take into account our data structure and sampling methods must be used, such as for example giving each infant a multi-dimensional matrix and clustering the matrices as a whole. See Tait and McNicholas (2019) for a potential approach.

5.1.2 Interpretation of factors

The features used in this thesis were carefully selected or created to reflect qualitative aspects of GMs that are common in GM evaluation. These aspects are the complexity, variability, and fluency of movements. Ideally a low dimensional projection of the feature matrix should be found that projects all the features on these three axis so that a visualization of complexity, variability, and fluency can be seen and infants can be rated along these "factors".

5.1.3 Clustering with abnormal GMs

The immediate next step is to collect GM data from babies who are considered at risk of having developmental disorders. These include babies who are born pre-maturely, or those who had complications during birth. Once the "abnormal" GMs are collected, they should be tested against the healthy GMs through training classifiers or comparing against healthy GM clusters.

As noted in the background (section 1.1), GM evaluation is being looked at for early evaluation of many pathologies including various forms of Cerebral Palsy, but also Autism Spectrum Disorder. It is important that more data is collected from at-risk infants and compared with the low-risk infant database in order to establish feature markers that distinguish the two groups clearly. This would also allow automation of diagnosis, which is one of the main desirable outcomes of this research direction.

5.1.4 Misclassified infants

An in-depth investigation into the infants who were significantly misclassified needs to be conducted as well. Though not many infants were significantly misclassified so as to cause concern for the integrity of the classifiers, but it is important that these misclassifications be investigated, through a non-black box model, to understand what aspects of their GMs caused them to be misclassified into significantly higher or lower age groups than their respective age group.

5.2 Summary of thesis

In this thesis, we studied the classification of infants into various age groups based on their movement characteristics. These movement characteristics were carefully chosen to model aspects of GMs which clinicians use to qualitatively evaluate GMs. We showed that complexity and variability features help to discriminate between the different age groups and that we can achieve good precision and accuracy on predicting the age group. On the other hand, fluency related features did not have a significant importance for classification and this is in agreement with what the GM literature says about fluency in the age range of the infants who participated in this study. Though the classifiers achieved good accuracy on the 2 class classification task, they showed notable overlap when finer binning of age as a label was used. This was addressed by instead clustering the infants in a semi-supervised fashion. The clusters reflected different age groupings that are reflective of a healthy developmental trajectory, but also showed the variance that exists in infants where some infants develop at a faster or slower pace than others within their age group.

Bibliography

Balasubramanian, S., Melendez-Calderon, A., Roby-Brami, A. *et al.* On the analysis of movement smoothness. *J NeuroEngineering Rehabil* **12,** 112 (2015). https://doi.org/10.1186/s12984-015-0090-9

Bishop, Christopher M. *Pattern recognition and machine learning*. springer, 2006.

Blauw-Hospers, C. H., et al. "Does early intervention in infants at high risk for a developmental motor disorder improve motor and cognitive development?." Neuroscience & Biobehavioral Reviews 31.8 (2007): 1201-1212.

Bosanquet, Margot, et al. "A systematic review of tests to predict cerebral palsy in young children." Developmental Medicine & Child Neurology 55.5 (2013): 418-426.

Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.

Cans, Christine. "Surveillance of cerebral palsy in Europe: a collaboration of cerebral palsy surveys and registers." Developmental Medicine & Child Neurology 42.12 (2000): 816-824.

Disselhorst-Klug, Catherine, et al. "Introduction of a method for quantitative evaluation of spontaneous motor activity development with age in infants." Experimental brain research 218.2 (2012): 305-313.

Einspieler, Christa, and Heinz FR Prechtl. "Prechtl's assessment of general movements: a diagnostic tool for the functional assessment of the young nervous system." Developmental Disabilities Research Reviews 11.1 (2005): 61-67.

Feller, Marla B. "Spontaneous correlated activity in developing neural circuits." Neuron 22.4 (1999): 653-656.

Ghahramani, Zoubin, and Geoffrey E. Hinton. *The EM algorithm for mixtures of factor analyzers*. Vol. 60. Technical Report CRG-TR-96-1, University of Toronto, 1996.

Gowda, Suraj, et al. "Accelerating submovement decomposition with search-space reduction heuristics." IEEE Transactions on Biomedical Engineering 62.10 (2015): 2508-2515.

Gravem, D., et al. "Assessment of infant movement with a compact wireless accelerometer system." Journal of Medical Devices 6.2 (2012): 021013.

Hadders-Algra, Mijna. "General movements: a window for early identification of children at high risk for developmental disorders." The Journal of pediatrics 145.2 (2004): S12-S18.

Hadders-Algra, Mijna. "Putative neural substrate of normal and abnormal general movements." Neuroscience & Biobehavioral Reviews 31.8 (2007): 1181-1190.

Hadders-Algra, Mijna. "Early diagnosis and early intervention in cerebral palsy." Frontiers in neurology 5 (2014): 185.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Hebbeler, Kathleen, et al. "Early intervention for infants and toddlers with disabilities and their families: Participants, services, and outcomes." Menlo Park, CA: SRI International (2007).

Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. "A practical guide to support vector classification." (2003): 1396-1400.

Kanemaru, Nao, et al. "Specific characteristics of spontaneous movements in preterm infants at term age are associated with developmental delays at age 3 years." Developmental Medicine & Child Neurology 55.8 (2013): 713-721.

Karch, Dominik, et al. "Kinematic assessment of stereotypy in spontaneous movements in infants." Gait & posture 36.2 (2012): 307-311.

Karpievitch, Yuliya V., et al. "An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF++." PloS one 4.9 (2009): e7087.

Kato, Moe, et al. "Decomposition of spontaneous movements of infants as combinations of limb synergies." Experimental brain research 232.9 (2014): 2919-2930.

Kaufman, Leonard, and Peter J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons, 2009.

Ma, Liang, et al. "Effect of early intervention on premature infants' general movements." Brain and Development 37.4 (2015): 387-393.

Marcroft, Claire, et al. "Movement recognition technology as a method of assessing spontaneous general movements in high risk infants." Frontiers in neurology 5 (2015): 284.

McLachlan, Geoffrey, and David Peel. "Mixtures of factor analyzers." *In Proceedings of the Seventeenth International Conference on Machine Learning*. 2000.

McNicholas, Paul David, and Thomas Brendan Murphy. "Parsimonious Gaussian mixture models." *Statistics and Computing* 18.3 (2008): 285-296.

McNicholas, Paul D., and Thomas Brendan Murphy. "Model-based clustering of microarray expression data via latent Gaussian mixture models." *Bioinformatics* 26.21 (2010): 2705-2712.

McNicholas, Paul David, Aisha ElSherbiny, Aaron F. McDaid and T. Brendan Murphy pgmm: Parsimonious Gaussian Mixture Models. R package version 1.2.4. https://CRAN.R-project.org/package=pgmm (2019)

Mosley, Lawrence, "A balanced approach to the multi-class imbalance problem" (2013). *Graduate Theses and Dissertations*. 13537. https://lib.dr.iastate.edu/etd/13537

Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." *the Journal of machine Learning research* 12 (2011): 2825-2830.

Prechtl, H. F. R. "Qualitative changes of spontaneous movements in fetus and preterm infant are a marker of neurological dysfunction." (1990): 151-158.

Prechtl, Heinz FR, et al. "An early marker for neurological deficits after perinatal brain lesions." The Lancet 349.9062 (1997): 1361-1363.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ (2018).

Steele, Russell J., and Adrian E. Raftery. "Performance of Bayesian model selection criteria for Gaussian mixture models." *Frontiers of statistical decision making and bayesian analysis* 2 (2010): 113-130.

Richman, JS; Moorman, JR (2000). "Physiological time-series analysis using approximate entropy and sample entropy". *American Journal of Physiology. Heart and Circulatory Physiology*. **278** (6): H2039–49

Shonkoff, Jack P., and Penny Hauser-Cram. "Early intervention for disabled infants and their families: a quantitative analysis." Pediatrics 80.5 (1987): 650-658.

Tait, Peter A., and Paul D. McNicholas. "Clustering higher order data: Finite mixtures of multidimensional arrays." *arXiv preprint arXiv:1907.08566* (2019).

Taga, Gentaro, Rieko Takaya, and Yukuo Konishi. "Analysis of general movements of infants towards understanding of developmental principle for motor control." Systems, Man, and Cybernetics, 1999. IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on. Vol. 5. IEEE, 1999.

Witten, Daniela M., and Robert Tibshirani. "A framework for feature selection in clustering." *Journal of the American Statistical Association* 105.490 (2010): 713-726.

Xu, Ji, Daniel J. Hsu, and Arian Maleki. "Benefits of over-parameterization with EM." *Advances in Neural Information Processing Systems*. 2018.

Zhang, Cha, and Yunqian Ma, eds. *Ensemble machine learning: methods and applications*. Springer Science & Business Media, 2012.