Classification of Fetal Heart Rates

Stony Brook University
Department of Electrical Engineering
ESE 441

Ian Jacobsen
William Dwyer

Abstract

A fetus suffering from oxygen insufficiency can face a number of challenges, such as neurodevelopmental disabilities, and in some cases death. It has been shown that the pH of the blood in the arterial cord contains valuable information regarding antenatal oxygen sufficiency [1]. Practitioners regularly use fetal heart rate (FHR) data to make predictions on the health of a fetus, but human observations of results are limited in accuracy and effectiveness. There are a number of invasive and non-invasive methods to retrieve FHR data. Fetal scalp blood testing is an invasive method that is used for monitoring fetal oxygen levels. Electronic fetal monitoring (EFM), or cardiotocography (CTG) is a common non-invasive method for monitoring the fetal heart rate in Western hospitals.

The main focus of this research is to implement generative models using Naïve Bayes and Hidden Markov Models (HMM) of both healthy and unhealthy fetuses as discriminated by their pH values and then using log likelihoods to classify unknown fetal heart rate data as either coming from either a healthy or unhealthy fetus. Using FHR data collected and compiled in an open access database courtesy of Goldberger, we perform a number of data cleaning procedures and then extract numerous features of the time-series data that hopefully provide information on the value of a fetus's pH. These features along with the fetuses' known health outcomes are what we train our generative models to. We aim to determine whether these methods are well suited to the need of fetal health classification and what parameters are best suited for the models constructed.

1. Introduction

Childbirth is an extremely important moment in a parent's life and arguably the most important day of the child's. Childbirth can often be an extremely stressful event for soon-to-be parents; the parent carrying the child often experience intense pain and physical distress along with the uncertainty over the health of the future child. The circumstances of the child in the womb can affect their development for the rest of their life, so it is crucial for doctors to be able to accurately assess the health of the fetus to determine what interventions may be required to care for it. Unfortunately, due to the fetus being in the parent's womb, we are limited in the ways we can assess the fetus's health.

Modern technology has developed many tools for gaining information on a fetus's development, from basic tools, such as stethoscopes, to more sophisticated hardware, like ultrasound. Fetal heart rate (FHR) used to be measured periodically using stethoscopes, but as technology has advanced we have gained more precise ways to measure FHR [1]. Electronic fetal monitoring (EFM) allows us to collect long term and more accurate information on a fetus's health. One form of EFM is Cardiotocography (CTG), which monitors the heart rate of the fetus. Previously doctors would have to visually observe the FHR and determine based on agreed upon guidelines how to interpret the results and assess the health of the fetus. While this was still a large advance in trying to improve birth outcomes, there were many factors that restricted the usefulness of FHR analysis. First, the data is often noisy. Due to the nature of fetal development, we do not have direct access to the fetus and therefore our data often has flaws. Second, following the same guidelines, professionals often have large discrepancies in their opinions of a fetus's health. In [1] for example, they find only about 50% inter-observer agreement [1]. Finally, there is only so much a human can discern by visually observing a FHR signal. Humans

do not have the ability to detect hidden mathematical relations that quantify information in the data and relate it to the health outcomes of fetuses.

Fortunately, there is growing hope in automatic analysis of FHR using signal processing and machine learning techniques. Computers are well equipped to do the detailed analysis of FHR that humans are not capable of. The goal of this project is to determine informative characteristics of FHR signals that are related to the health outcome of fetuses. Using different linear and non-linear features of the FHR signal we aim to build generative models that represent processes that might reflect the generative process of given FHR signals. The two generative models tested in our work are Naïve Bayes and Hidden Markov Models (HMM) with the goal of seeing how these techniques can contribute to the prediction of fetal health states.

2. Background

2.1 Survey

The field of automatic analysis of FHR has recently been gaining attention from researchers of a diverse range of fields. There have been many methods of processing data, representing classes of fetal health outcomes, and features that are used to determine the state and health of a fetus. [1] provides an overview of the current research that we have benefitted greatly from.

A major source of motivation for researching FHR analysis is to detect prolonged moments of anaerobic metabolic activity that can lead to hypoxanemia, hypoxia, and asphyxia. These conditions are states of oxygen deficiency of increasing severity [1]. If the fetus remains in these states for too long, it can suffer serious long-term damage that can affect the life of the

fetus [1]. A healthy fetus should be able to respond to these conditions, and hopefully remedy the oxygen insufficiency through a physiological process that will result in changes in the FHR. The fetus's response to these conditions results in a change in the FHR pattern [1]. Other fetal activity and activity of the carrying parent can also contribute to FHR changes [1]. There are two primary methods for obtaining the FHR: using an ultrasound probe and transducer before uterine membrane rupture and attaching an electrode directly to the fetuses scalp after membrane rupture [1]. Generally the direct method of measurement is more accurate while the external method is beneficial due to it being non-invasive and easy to use [1].

A commonly used biological marker for determining the health outcome of a fetus is its blood pH level [1]. Generally, a higher pH score is indicative of a healthy fetus while a lower score is related to negative health outcomes [1]. Previous works have primarily focused on a binary classification between healthy and unhealthy fetuses with thresholds ranging from 7.00-7.17 [1]. There is still debate over what the proper biological marker is for fetal condition, though generally pH is considered the most robust [1], [2]. Before automatically analyzing the FHR, it should be cleaned for an accurate assessment and computation of features. FHR data often has a large amount of artifacts due to movements of the parent and fetus, therefore these should be detected and dealt with before using the data [1]. [1] suggests a method that detects difference in successive values that is greater than 25 bpm and replaces them using linear interpolation as long as the length of missing data was less than 15 seconds.

Various features have been proposed in the analysis of FHR data that range from easily detectable by human visual analysis to requiring complex computation. There are a few standard features that are used in current International Federation of Gynecology and Obstetrics (FIGO) guidelines which are a signal's baseline, accelerations, and decelerations [1]. Another widely

used feature is a signal's variability which is roughly how much a heart rate changes in a specific amount of time. There exist two main kinds of variability, short and long term which measure the amount of change in a signal over short and long periods of times respectively [1]. Yeh et al developed a measure called Interval Index (II) which characterizes low-frequency irregularity of a signal [3]. This has been implemented by relating a measure of short term variability (STV), beat-to-beat variability, to the standard deviation (std) of a signal [4], [5]. Frequency features have also been used and usually break the signal into different spectral bands that represent different fetal and parental bodily activity [6]. In addition to standard linear features, many non-linear features have been proposed. These include fractal dimension analysis, detrend fluctuation analysis, entropy, Lempel Ziv complexity, and Poincaré plots [1], [7]–[10].

After selecting features that can be extracted from the FHR, and possible output classes, the goal is to determine methods of pattern recognition and machine learning that can be used to create a method for predicting the health of future fetuses. There have been many previously used and proposed methods with varying degrees of success. Most previous examples have focused on a binary classification model of healthy and unhealthy fetuses using various pH thresholds as cited above as cutoffs for a given class. Simple classification techniques include k-nearest neighbors and linear discrimination. In Gonçalves et al, data was separated using a simple linear discriminator and achieved promising results using just approximate entropy (ApEn) and II [2]. Other more in-depth learning methods that have been used for this task include artificial neural networks, hidden Markov models, and support vector machines. In addition to simply creating classifiers, parameters are often varied and tested using validation methods such as k-fold cross-validation (CV) to confirm that the classification model is the

optimal under certain constraints. CV was used in the work of Spilka [9] and leave-one-out was used by Goncalves et al [2].

Previous work has shown the possible success of each method and provided inspiration for what to explore in our research. We plan on using the features previously developed by other researchers to train generative classification models including Naïve Bayes and Hidden Markov Models (HMM).

2.2 Planning

This project requires a diverse skill set drawing from a number of disciplines including signal processing, machine learning, mathematics and even some understanding of biological information. Since our goal is to develop a practical method of fetal health detection, we need to figure out how to process, analyze, and predict based on the information with nearly real-time processing. Due to the complex nature of the project, we are required to build upon our previously learned skills through research and experience. Many of the skills needed for this project lie in the intersection of electrical engineering and other fields, which means that we are required to broaden our areas of knowledge and engage in self-directed learning.

Our project is focused on the complex task of using noisy data to predict the class of data, with high-stakes consequences. This requires a wide range of background information. First is a thorough understanding of the domain of FHR analysis and what further knowledge will be required to begin making advances in the field. The primary source of guidance in this project is Jiří Spilka's dissertation "Complex Approach to Fetal Heart Rate Analysis: A Hierarchical Classification Model" [1]. It provides an in-depth look at the background of FHR analysis and the previous and current research in the automatic analysis of FHR. Spilka provides a brief

summary of relevant aspects of fetal development and physiology and the ways to determine and assess the health of a fetus. He also overviews the ways in which FHR is related to different fetal conditions. It also highlights the many challenges involved in this field and how he and others have suggested addressing them. The length and breadth of the dissertation acts as a good introduction and overview to the relevant topics and knowledge needed to make progress in the field. It also cites numerous papers and sources of external tools to delve deeper into the material.

Signal processing skills are important when working with noisy digital data. Electrical engineering coursework, such as “ESE 305 deterministic signals and systems”, “ESE 306 random signals and systems”, and “ESE 337 digital processing: theory”, has provided a background in many of the skills required to filter and manipulate FHR data. Our professor also provided us with various techniques to enhance data such as interpolation and filtering techniques. Papers and online tools will also be used to further our understanding of this aspect of design. In addition to processing the data to be useable, we also need to decide on how to use our data. Our project is focused on associating FHR data with fetal health. This can be seen as a problem of pattern recognition and machine learning. These multi-disciplinary fields require knowledge of engineering, computer science and mathematics to determine how to best represent our problems by selecting features and classes that represent the factors of our project. An introduction to many topics in the fields of pattern recognition and machine learning was provided in elective coursework such as “ESE 358 computer vision”, “CSE 353 machine learning”, and “CSE 352 artificial intelligence”. Some skills under this domain include how to determine and select informative features from FHR data, how to select output classes that represent the health of the fetus, when to use continuous or discrete representations of inputs and

outputs, and different techniques for creating classifiers and learning on data. General background for this domain will be learned from various papers online as well as massively open online courses (MOOCs). Spilka's dissertation provides a good background of previously used techniques for creating features out of FHR data along with references to other papers and tools for further understanding of the material.

Our current plan is to implement our project using MathWorks MATLAB programming environment. This means we will need to learn how to implement signal processing, pattern recognition, and machine learning techniques in the MATLAB language. For this online resources such as MathWorks MATLAB documentation will be an invaluable tool for learning what is needed to implement our project.

3. Original Work

3.1 Design Contstraints

Our project forces us to consider a number of constraints when forming a plan for analyzing FHR. Financial constraints are negligible due to the free MATLAB license provided by Stony Brook University, and there are no foreseen additional costs. The biggest constraints that we face are technical and ethical constraints. When working with human health data there will always be a need to consider how to responsibly obtain and use the data and technical constraints are inevitable in any project.

Some practical constraints that have to be considered involve our access to data and quality of the data that is available. We have access to a fairly large database of FHR data from the Czech Technical University – University Hospital Brno that contains 552 samples that was

developed by Spilka and described in his thesis [1]. Each sample is missing 40% or less data, though this still leaves a risk of affecting our calculated features [1]. This also means that the method of interpolation can strongly affect our outcomes. There is also a question of how much "good" data is affected by noise. Unfortunately with real world data, especially in a difficult to obtain situation like fetal development, noise and imperfect data must always be considered. Other problems come from the fact that using a threshold of 7.15 pH, ~80% of FHR samples are considered "healthy" which leaves only ~20% "unhealthy". This imbalance will affect the learning process and must be dealt with accordingly to properly achieve significant, reliable classification. The data was collected and anonymized with informed consent by the women carrying the fetuses and approved for collection by the Institutional Review Board of University Hospital Brno [1].

While right now our project is primarily calculating and testing various feature sets to see which metrics contain the most relevant information, it is important to consider the future clinical application of our work, which imposes further constraints. The ultimate goal is to create a system that could, in real time, detect the health of a fetus and determine if action needs to be taken to protect the life and future development of the fetus. This means the system must be computationally efficient to produce results quickly and that they must be accurate, considering lives are dependent on its success. It must also be able to deal with a fair amount of noise considering we don't have the luxury of hand-selecting clean data samples in the field.

3.2 Selected Features

Our goal in feature selection was to determine quantifications of features of FHR signals that provide information about the health state of a fetus. This section provides a description of

some of the features tested in our work. One class of features that we used were standard time domain features commonly used in FHR analysis. We calculated simple features such as the mean, median, standard deviation and median deviation. A commonly used statistic is short term variability (STV) which finds the average change in FHR period. Our implementation of STV is drawn from [4]–[6]. We define STV by:

$$STV = \frac{1}{24M}\sum_{i=1}^{24M}|sm(i+1) - sm(i)|,$$

where sm(i) is the value of our signal on 2.5s intervals and M is the length of the signal in minutes [4], [6]. Alternative quantifications of short term variability are statistics known as short term irregularity (STI) and modified STI which are defined as follows:

$$IQR(tan^{-1}(s(i+1)/s(i))) ,$$

$$IQR(s(i+1) - s(i)),$$

where s(i) is the FHR signal at a given time step i [11].

In addition to how the FHR signal may vary overtime, we also quantify the signals long-term variability (LTV). The first LTV statistic is the delta (Δ) value which is given by:

$$\Delta = \frac{1}{M}\sum_{m=1}^{M}\left|\max_{i\in m}\big(s(i)\big) - \min_{i\in m}\big(s(i)\big)\right|$$

with max and min computed within each minute of the signal and M being the number of minutes in the signal [1]. The second LTV statistic is long term irregularity (LTI) defined as:

$$IQR(\sqrt{s(i+1)^2 - s(i)^2})$$ [11].

Another time domain statistic is the Interval Index (II) first proposed in [3]to describe low frequency irregularity of heart rate signals. We used the implementation in [12]which is defined for one minute of a signal by:

$$II = \frac{std[sm(i+1) - sm(i)]}{STV}, i = 1, \dots, 23,$$

where std is the standard deviation of the difference of beat values.

In addition to time domain features, we also analyzed the spectral content of the signals. We divided the frequencies in the signal into four regions very low frequencies (VLF): $f_0 - f_{VLF}$, low frequencies (LF): $f_{VLF} - f_{LF}$, medium frequencies (MF): $f_{LF} - f_{MF}$, and high frequencies (HF): $f_{MF} - f_{HF}$. We chose $f_{VLF} = 0.06$, $f_{LF} = 0.3$, $f_{MF} = 1$, $f_{HF} = 2 = fs/2$ as suggested by [12]. We also analyzed the LF/(MF + HF) ratio as recommended in [5], [13]. These frequency bands are thought to correlate to various bodily activities of the fetus and the carrying parent [5]. The LF range is thought to quantify sympathetic control and vasomotor activity while the HF range is correlated with respiratory activity [5]. The MF range is thought to relate to physical movement by the fetus and the carrying parent [5]. It is suggested that the LF/(MF+HF) ratio is related to the balance of the autonomic nervous system activity [12].

We also utilized non-linear features. One aspect of a signal is the unpredictability of new information, which is related to its regularity. This was formalized as the concept of entropy first introduced in the work of Claude E. Shannon [14]. An example of the lowest possible entropy, zero, would be if a string of all 1s. In this example, you would always know what value would come next. A high entropy system would be the flipping of a perfectly weighted coin so that the probability of heads and tails are exactly equal. Entropy would be a useful feature to distinguish

one signal from another signals. Unfortunately, you cannot fully determine the entropy of a signal of finite length.

To address this problem approximate entropy was developed. Approximate entropy (ApEn) was put forward by Pincus from the idea of Kolmogorov-Sinai entropy [8]. ApEn is developed by creating vectors of length *m* from the FHR data. For each data point a vector of length *m* is compared to all other vectors in the data sample. $n_i^m(r)$ is the number of vectors that are with a distance less than *r*, which is a parameter that is usually set at a value of .15 or .2 times the standard deviation of the data [1], [9]. Any distance function can be used but we used the Chebyshev distance as this was used in Richman and Moorman's paper [9].

$$n_i^m(r) = \#\ of\ vectors\ where\ \{d[x_m(i), x_m(j)] < r : 1 \leq j \leq N - m\}$$

$$C_i^m(r) = n_i^m/(N - m + 1)$$

$$\Phi^m(r) = \frac{1}{(N - m + 1)} \sum_{i=1}^{N-m+1} \ln C_i^m(r)$$

$$ApEn\ (m, r) = \Phi^m(r) - \Phi^{m+1}(r)$$

While ApEn is still commonly used as a measure of entropy and a feature in heart rate analysis, it has some drawbacks. First, it is a biased statistic. When calculating $C_i^m(r)$ self matches are included to ensure that the natural logarithm of each term is guaranteed to be defined. This leads to a measure of entropy consistently lower than the expected entropy. Approximate entropy is also largely dependent on sample size. To account for some of these issues, Richman and Moorman have developed this work to create Sample Entropy (SampEn)

[Richman]. SampEn draws from the same theoretical framework as ApEn but through a slightly modified formulation. For a template vector starting at data point $x(i)$, SampEn defines two values, $B_i^m(r)$ and $A_i^m(r)$ to represent the probability of a match for a given template vector of length m and $m+1$ respectively. These probabilities are calculated by finding all vectors where $d < r$ not including self-matches and dividing by the number of vectors, with d again being Chebyshev distance. There is a $B_i^m(r)$ and $A_i^m(r)$ value for every vector and these values are then averaged giving us the terms $B^m(r)$ and $A^m(r)$. SampEn is than defined as the negative natural logarithm of $A^m(r)$ over $B^m(r)$ as shown in the equations below.

$$n_i^m(r) = \#\ of\ vectors\ where\ \{d[x_m(i), x_m(j)] < r : 1 \le j \le N-m\ \&\ j \ne i\}$$

$$B_i^m(r) = n_i^m/(N-m-1)$$

$$A_i^m(r) = n_i^{m+1}/(N-m-1)$$

$$B^m(r) = \frac{1}{N-m}\sum_{i=1}^{N-m} B_i^m(r)$$

$$A^m(r) = \frac{1}{N-m}\sum_{i=1}^{N-m} A_i^m(r)$$

$$SampEn = -\ln\frac{A^m(r)}{B^m(r)}$$

A third representation of entropy is fuzzy entropy that can be represented by the following equations:

$$X_i^m = \{s(i), s(i+1),\ \ldots,\ s(i+m-1)\} - s0(i)$$

$$s0(i) = \sum_{k=0}^{m-1} s(i+k)$$

$$d_{ij}^m = \max_{k\in(0,m-1)} |[s(i+k) - s0(i)] - [s(j+k) - s0(j)]|$$

$$D_{ij}^m(n,r) = \mu\left(d_{ij}^m, n, r\right) = e^{-\frac{(d_{ij}^m)^n}{r}}$$

$$\phi^{\,m}(n,r) = \frac{1}{N-m}\sum_{i=1}^{N-m}\left(\frac{1}{N-m-1}\right)\sum_{j=1,j\neq 1}^{N-m} D_{ij}^m$$

$$FuzzyEn(m,n,r) = \ln \phi^{\,m}(n,r) - \ln \phi^{\,m+1}(n,r)$$ [10].

Another non-linear feature used is Higuchi's dimension, first proposed in [15]. This feature is meant to describe the irregularity of a time-series [15]. Given a time-series s(i), we construct subseries as follows:

$$s_k^h: s(h),\ s(h+k),\ s(h+2k) \ldots, s\left(h + \left\lfloor\frac{N-h}{k}\right\rfloor k\right)\ \ h = 1,\ 2, \ldots, k$$

where m and k are integers that indicate the initial time of the series and the time interval, respectively. The length of the curve s_k^h is then:

$$L_h(k) = \left\{\left(\sum_{i=1}^{\left\lfloor\frac{N-h}{k}\right\rfloor} |s(h+ik) - s(h+(i-1)k)|\right)\frac{N-1}{\left\lfloor\frac{N-h}{k}\right\rfloor k}\right\}/k$$

The length of the curve for a time interval k ($\langle L_h(k)\rangle$) is defined as the average value over k sets of $L_h(k)$ [15]. If $L_h(k) \propto k^{-D}$ then the Higuchi's dimension is D.

Poincaré plots allow for the visualization of heart rate variability and nonlinear aspects of the FHR [7]. We used three descriptors of the plot, the first two are defined as follows:

$$SD1 = \sqrt{\frac{1}{2}\sigma_{SD}^2}$$

$$SD2 = \sqrt{\frac{1}{2}\sigma_{FHR}^2 - \frac{1}{2}\sigma_{SD}^2}$$

and measure the gross variability of the FHR [7]. The third descriptor is Complex Correlation Measure (*CCM*) which can be approximated as a function of the autocorrelation of the series with different amounts of lag as follows:

$$\mathrm{CCM}(m) = F[r_s(m-2), r_s(m-1), r_s(m+1), r_s(m+2)]$$

with $r_s(*)$ being the autocorrelation function [7].

3.3 Procedure

3.3.1 Preprocessing and Feature Extraction

The first step in performing the analysis on the aggregated fetal heart rate data is to remove artifacts that were introduced into the raw data by means of human or machine error. Recording data with a CTG is a very sensitive process that is prone to falsely appending erroneous data. The most common cause of this is due to movements of the mother or the fetus [1]. Due to the high sensitivity of CTG, preprocessing on the raw data must be performed before any analysis may take place. Extreme outliers must be removed, and periods of missing data must be remedied by artificially appending data produced by means of interpolation. This

preprocessed data is free of obvious irregularities, which can be confirmed by constructing a time-plot and observing that there are no extreme (greater than 25 bpm) jumps in FHR [6].

As with any time-series, it can be assumed that the FHR is composed of trending, and random components. The random component contains most of the information of the signal, and therefore this component is of most interest for analysis, although the trending component may also contain important information. There are many published methods for isolating the random component of a time-series. In our analysis we have experimented with three different methods for isolating the random component. Two of the methods are used to estimate the baseline of the FHR, from which we subtract from the original data to be left with the morphological component. These two methods are referred to as a moving average filter, and a moving median filter. The third method is a differencing approach, which directly gives the morphological component.

$$movingAv(k) = \frac{1}{2q+1}\sum_{i=-q}^{q} x(k+i)$$

$$movingMed(k) = median\{[x(k-q), x(k+q)]\}$$

$$diff(k) = (1-B)x(k) = x(k) - x(k-1)$$

A selection must be made for the window size, denoted $2q+1$, of the moving average and moving median filters. We have found that a symmetric noncausal window of size 941 samples (approximately 3.9 minutes with a sampling frequency of 4Hz) results in a fair balance of smoothness and retained information. A routine was written in MATLAB which uses the preprocessed FHR data to construct and return a vector containing the morphological component of the final 30 minutes of collected data.

Features must be derived from the CTG data. The goal is to derive as many uniquely informative features as possible. Common quantities that are used by medical professionals were computed, as well as additional features that are seldom used. Features such as Approximate Entropy, Sample Entropy, Interval Index, Short-Term-Variability, Long-Term-Irregularity, mean, and standard deviation were calculated over the final 29 minutes. Due to excessive noise and missing data during the final moments before birth, the last four minutes of each FHR time series was truncated. The final step in the preprocessing was to visually scan each raw-FHR time series and remove the training examples that had an excessive amount of missing data.

A set of 400 training examples was chosen to be used for learning, consisting of 350 healthy fetuses (pH > 7.15), and approximately 50 non-healthy fetuses (pH < 7.15). Once the features of each training example were obtained, they were regularized. This step is necessary because the order of magnitude of the raw calculated features vary from feature to feature. The most considerable difference in range is between Sample Entropy and Long-Term-Irregularity, varying by approximately two orders of magnitude. We used two methods of feature regularization as follows:

$$x^{j}_{reg0}(i) = \frac{x^{j}(i) - x^{j}_{min}}{x^{j}_{max} - x^{j}_{min}}$$

$$x^{j}_{reg1}(i) = \frac{x^{j}(i) - \bar{x}^{j}}{\sigma}$$

After the regularization, principle component analysis was performed on the set of training features. The purpose of the PCA is to reduce the dimensionality of the feature matrix by forming an uncorrelated orthonormal basis that is capable of accurately representing the original feature space with minimal loss of information.

3.3.2 Classification

3.3.2.1 Naïve Bayes Classifier

A Naïve Bayes classifier is a generative classification technique that was used in order to classify a fetus into one of two classes; healthy or unhealthy. The key assumption to make when using the Naïve Bayes classifier is that each observation is drawn from a multivariate Gaussian distribution, and that each observation in time is independent of every other observation in time. We estimate the parameters of the multivariate Gaussian using the sample mean and sample covariance estimators for each observation in time. Therefore, the emission probability of one time-sample given the class that it belongs to is described by the equation:

$$p(x|class) = \frac{e^{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^{\mathrm{T}}C^{-1}(\boldsymbol{X}-\boldsymbol{\mu})}}{(2\pi)^{n/2}|\boldsymbol{C}|^{\frac{1}{2}}}$$

where μ is the mean vector that contains the mean for each feature and C is the covariance matrix.

Two models are formed, each having a time series of mean vectors and covariance matrices. Once the models are trained, they can be used to classify new fetuses. The classification is performed by maximizing the likelihood that the fetus in question belongs to each of the classes. In order to prevent underflow, the log-likelihood was used, and therefore the likelihood measurement was obtained by summing the log-likelihoods of the fetus belonging to each class over each sequence of observations.

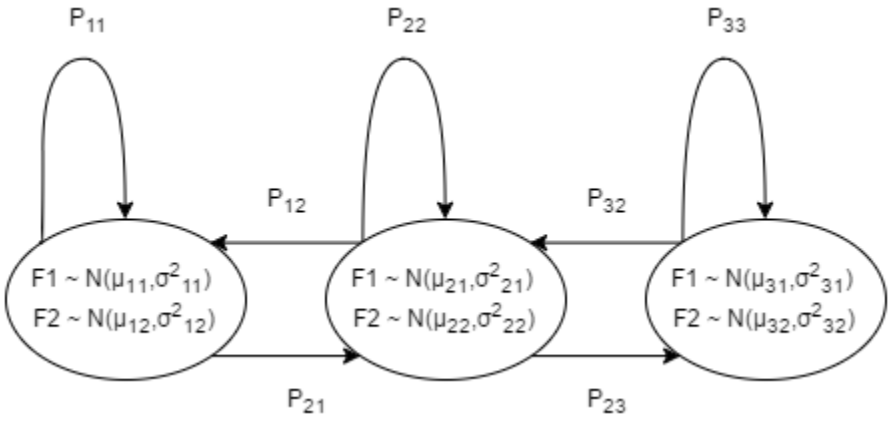
3.3.2.2 Hidden Markov Models

A Hidden Markov Model is a specific type of Bayesian network designed to represent probability densities of observed variables [16]. A time-series is divided into even duration

samples or observations [16]. The general concept of the HMM model is that observations are dependent on the state S_t of some stochastic process with a finite state space that is hidden to the observer [16]. The HMM model assumes that each state, S_t, is dependent only upon the previous state of the process, S_{t-1}, and that any observation at time t, Y_t, only depends on the state at of the process at the time, S_t [16]. These assumptions allow us to represent the probability of a given state and observation sequence with the following formula:

$$P(S_{1:T}, Y_{1:T}) = P(S_1)P(Y_1|S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})P(Y_t|S_t),$$

where $S_{1:T}$, $Y_{1:T}$ represent all the states and observations from time 1 to time T. The probability that a given observation will occur in a current state, $P(Y_i|S_j)$, can be modelled by a number of different probability functions [16].



Generic Markov Model Visualization with 3 states and 2 Gaussian observation features

In our work, we constructed two HMM models, one for healthy fetuses and another for unhealthy fetuses under the assumption that the time-series of the two classes would differ in their Markov processes. We assumed a mixture-of-Gaussians distribution for the probability a

given observation will occur in a given state, $P(Y_i|S_j)$. The number of mixtures (M) and states (Q) of our proposed model can be varied and evaluated based on cross-validated performance of each model. Before the HMM models can be used to classify unknown FHR data, they must first be constructed by inferring the transition matrix, prior state distribution, emission parameters (mean 'μ' and covariance 'σ^2') of training examples using forward-back algorithms and expectation maximization. After constructing the two models, future FHR samples are compared to these models to find the likelihood that it belongs to the two models. Whichever model it has a higher likelihood of belonging to is how the test sample is classified. We used a tool by Kevin Murphy (https://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html) that can do inference and classification of HMM models.

In order to test the accuracy of our models and to select the best model parameters we used a method called k-fold cross-validation. The essential idea is to get a better indication of the error in our models than if we only used fixed training data. The first step of k-fold cross validation is to divide data into k random subsets (folds) of equal size. In order to make sure each fold is representative of the full data set, we made sure the ratio of healthy to unhealthy fetuses was the same as for the total data set (approximately 80% healthy, 20% unhealthy) although the data is still randomly selected to reduce biasing an individual fold. After splitting the data into k sets, we do the cross-validation step. This is an iterative process which repeats k times. For each fold, we separate that fold out as test data and use the rest as training data for our models. We then find the predictive performance of that iteration in terms of true-positives, true-negatives, and percent properly classified. After all k iterations, these values are averaged across all iterations to approximate the predictive power of the model with the given parameters.

The cross-validation procedure is repeated for different configurations of our learning methods to perform model selection.

3.4 Results

All results described are the averages after a 10-fold cross-validation. Naïve bayes did not yield very good results suggesting that this method may not be well suited for the classification of fetal health or that we need a better method of taking into account the differences between the skewed sets of data.

All Features PCA, Reg 1		Predicted	
		Healthy	Unhealthy
Actual	Healthy	92.8%	
	Unhealthy		28.9%

All Features PCA, Reg 0		Predicted	
		Healthy	Unhealthy
Actual	Healthy	91.9%	
	Unhealthy		32.2%

Poincare, Fuzzy, II, STV, STD No PCA		Predicted	
		Healthy	Unhealthy
Actual	Healthy	90.3%	
	Unhealthy		50%

Poincare, Fuzzy, II, STV No PCA		Predicted	
		Healthy	Unhealthy
Actual	Healthy	89.7%	
	Unhealthy		42.6%

For HMM, the first comparison we had was between our two types of data regulation. It's clear that the way we regularized the features had a big difference in performance of our classifier. This may be due to the large difference in the amount of healthy and unhealthy samples causing a bias towards the healthy range.

All Features PCA, Reg 1, 15 States, 3 Mixtures		Predicted	
		Healthy	Unhealthy
Actual	Healthy	73.1%	
	Unhealthy		43.3%

All Features PCA, Reg 0, 15 States, 3 Mixtures		Predicted	
		Healthy	Unhealthy
Actual	Healthy	95.5%	
	Unhealthy		5.6%

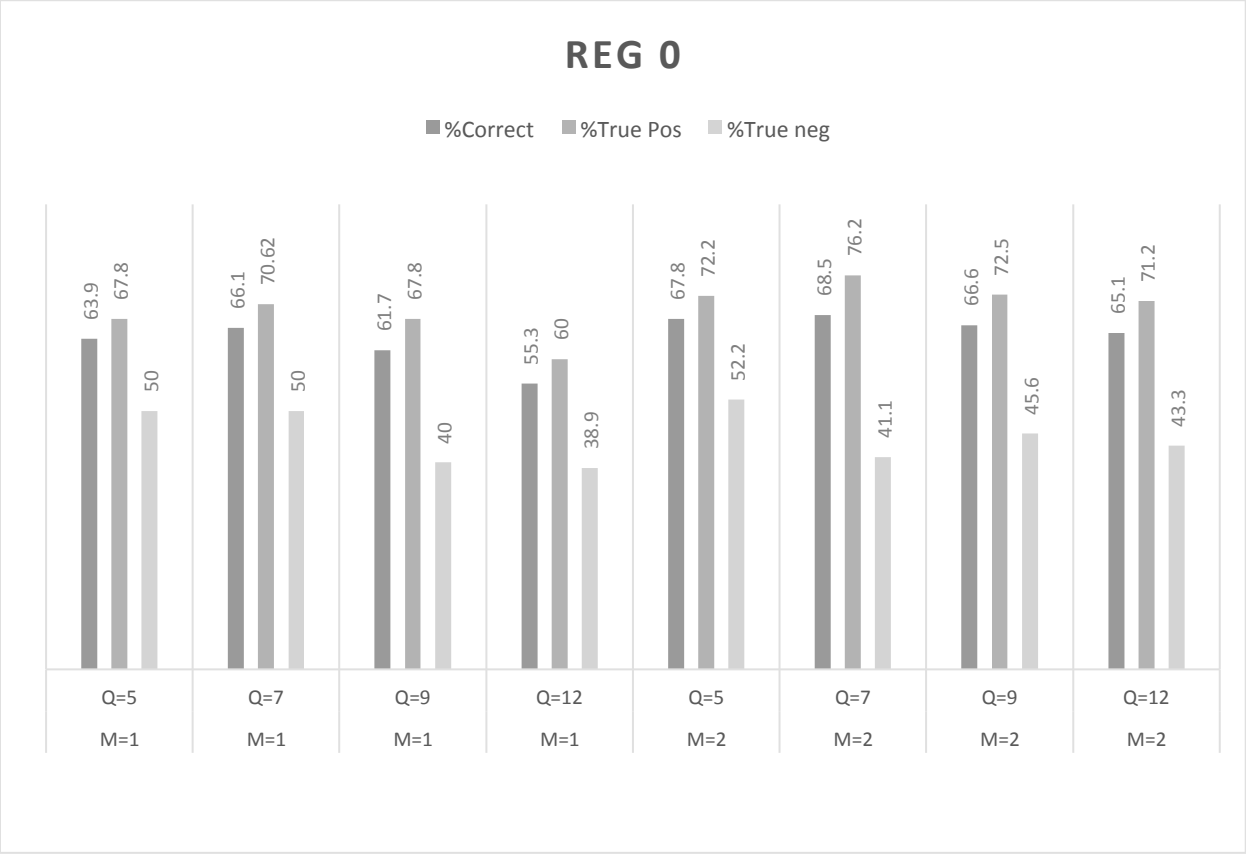
We did a few trials using limited features and yielded interesting results. When STV was added, unhealthy classification increased, but healthy classification plummeted in accuracy.

Poincare, Fuzzy Entropy, Interval Index, No PCA, 15 States, 3 Mixtures		Predicted	
		Healthy	Unhealthy
Actual	Healthy	90%	
	Unhealthy		33%

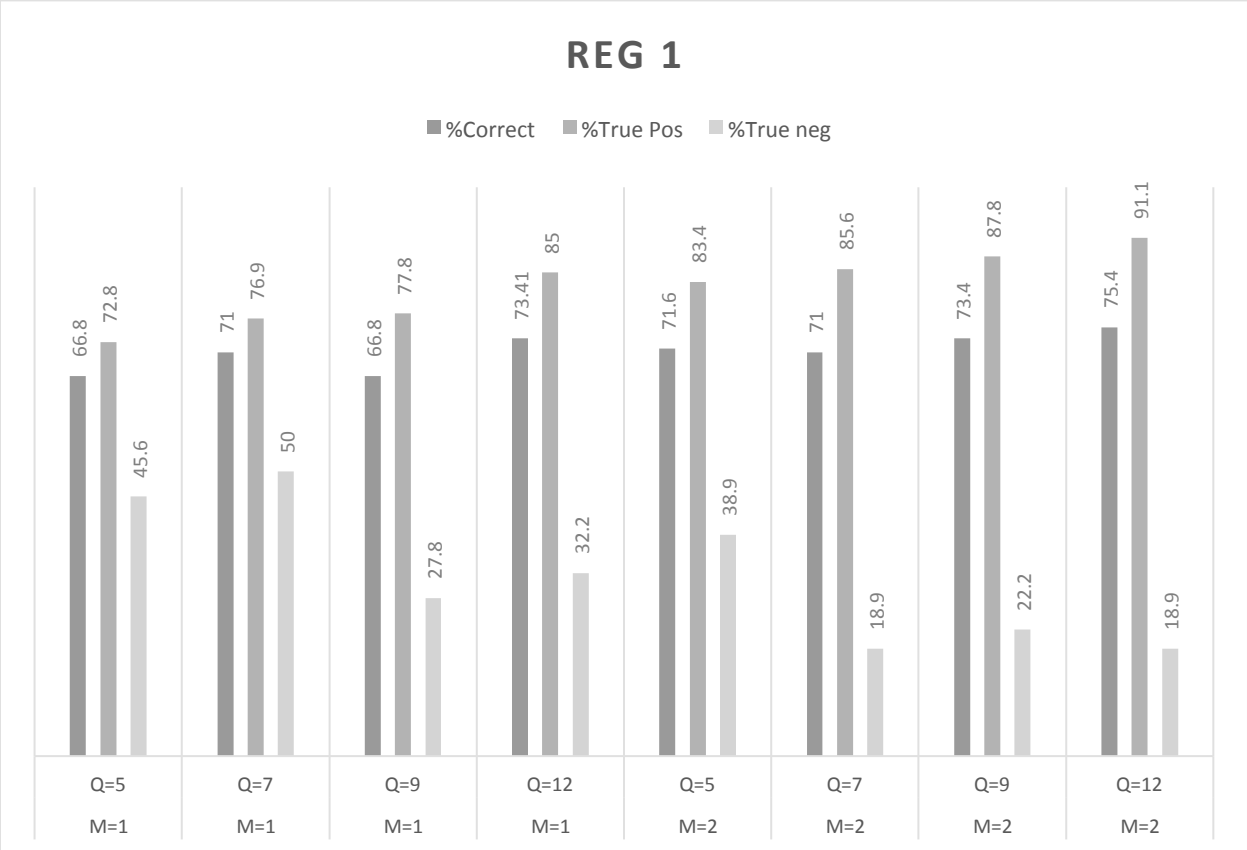
Poincare, Fuzzy, II, STV No PCA, 12 States, 3 Mixtures		Predicted	
		Healthy	Unhealthy
Actual	Healthy	65.6%	
	Unhealthy		52.6%

Next we focused on HMM with all features and PCA while varying parameters such as the number of states and number of Gaussians in the Mixture of Gaussians for both regularization methods. Looking first at the regularization method 0, we see that true negative

rates are consistently low, especially when compared to the relatively high true positive rate. Under both mixtures, we see the best true positive rate occurring with 7 states. When M is increased from 1 to 2, both true positive and true negative rates increase, suggesting that a single Gaussian output may not be enough to properly model the FHR signals. The best true negative rate is achieved when Q = 5 and M=2 as well as the third best true positive rate. This seems to achieve the optimal balance between healthy and unhealthy classification rates but still yields lower unhealthy classification rate than would be desired.



When we look at the same variations under regulation 1, we see drastically different results consistent with our other findings. True positive rates are consistently higher, but true negative rates suffer drastically. Both increasing Q and M appear to decrease the number of true negatives, while increasing true positive rates. For the our purposes, it appears the regulation format 1 is not well suited.



4. Discussion

4.1 Multi-disciplinary Experience/Issues

Our project depends upon multi-disciplinary knowledge and an ability to synthesize this information to yield successful results. FHR analysis draws upon knowledge from the domains of biology, computer science, engineering, and mathematics. There are different aspects of this project that are more suited to some fields than others but all are necessary for real progress. A biologist may try to relate known biological functions to different features to propose a biologically plausible model for classification, while an engineering approach may focus more towards models that function well. Both approaches can benefit from the work of the other. Theoretical mathematical concepts are often modified to work practically finding new methods of analysis. While each domain has its own focus, it is required that all researchers working on this field are required to know a little bit of each aspect in order to optimize results. Our team's broad background in electrical engineering, computer science, and applied mathematics and statistics gives us many areas to draw from and allows us to think about problems in many different ways.

4.2 Impact on Environment/Society

The ability to analyze FHR data is a significant achievement that would greatly impact society. FHR analysis and being able to predict the health of a fetus would greatly affect the field of obstetrics, protect the future life of fetuses, and prevent the distress to parents of a failed pregnancy to harm to their future child. The current guidelines for detecting changes in fetal health by analyzing FHR are not very precise, with only 50% inter-observer agreement as stated earlier in our paper [1]. A successful classification system would be much more precise and

accurate in predicting health states of a fetus, giving doctors the knowledge to know when to take action. This in turn would provide better care to the fetus and more peace of mind to parents. In addition to helping close to the time of delivery, further research into fetal analysis could lead to better monitoring of the fetuses health over longer periods to better manage care throughout a pregnancy. FHR research could also lead to a better understanding of the underlying conditions a fetus undergoes during development, sparking new research and biological discoveries.

4.3 Discussion of Results

The results we obtained through our research raise a number of interesting questions and point towards interesting future areas of exploration. The method of Naïve Bayes did not perform very well on the data we used which could be caused by a number of factors. The uneven amount of data may have been too much of an influence to learn each class properly, even though we tested priors to attempt accounting for this. It is also possible that the classes are not easily linearly separable, which would drastically affect the performance of the classifier. Considering the simplicity of the model and the more interesting and promising alternative of HMMs, Naïve Bayes is probably not an ideal method to pursue in future work.

Our results with HMMs suggested definite promise in the classification and understanding of healthy and unhealthy fetal heart rate patterns. With Q = 5, M = 2 and PCA of all features under regularization type 0, we obtained a classifier that was better than random selection for both healthy and unhealthy examples. While the rates achieved were not very impressive, they suggest the classifier is utilizing the information of the features to construct preliminary models. There is a general trend in our data that as true negative rates increase, true

positive rates decrease. This could suggest that the models we have constructed are very close to each-other making it hard to discriminate between the two classes. Something to consider in future work is whether the fundamental concept of their only being 2 classes that cause similar FHR responses is problematic and that creating only 2 models is where our model lies. This would account for why added mixtures increased true negatives; it could be accounting for greatly differing modes of fetal distress. Varying the pH threshold or creating more pH class ranges may help in better understanding and classifying FHR data.

5. Conclusion

Generative models for the classification of FHR data has been shown to be an area of study worth looking more into. While our preliminary results are still far from meeting a clinical level of accuracy, if can inform future work that will build on the foundation set in this paper. Further modification of parameters and possibly adding further state dependence would show if this is a good direction to be moving in, or if we will begin to reach a limit of how affective generative based models can be in the analysis and classification of FHR data.

6. Acknowledgements

This work was made possible by the guidance of Professor Petar Djurić who proposed the project and has led us through every step of the learning process. He has also provided us with valuable resources to learn more about research in the field. Kezi Yu and Inigo Urteaga have also been extremely valuable collaborators, helping us understand the data provided along with providing their insight to help correct and improve our work. We would also like to thank

Malvin de Núñez Estevez and Josue Nassar for discussions regarding our parallel research. Regular meetings with them have directed our efforts and introduced us to new ideas.

Thank you to Professor Wendy Tang and Tatiana Tchoubar for providing us with guidance in ESE 440/441 lecture and through references posted online. These have guided us through our project report and with our professional development. We also would like to acknowledge Stony Brook University, the College of Engineering and Applied Science, and the Department of Electrical and Computer Engineering for providing us with the education and resources to pursue meaningful work.

This work can only be done due to the work of previous researchers, especially Jiří Spilka whose dissertation has informed much of our work so far.

7. References

[1] J. Spilka, "Complex approach to fetal heart rate analysis: A hierarchical classification model," Czech Technical University in Prague, Czech Republic, 2013.

[2] A. Georgieva, S. J. Payne, M. Moulden, and C. W. G. Redman, "Artificial neural networks applied to fetal monitoring in labour," *Neural Comput. Appl.*, vol. 22, no. 1, pp. 85–93, 2013.

[3] S.-Y. Yeh, A. Forsythe, and E. H. Hon, "Quantification of Fetal Heart Beat-to-Beat Interval Differences.," *Obstet. {&} Gynecol.*, vol. 41, no. 3, 1973.

[4] G. Magenes, M. G. Signorini, and D. Arduini, "Classification of cardiotocographic records by neural networks," *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, vol. 3. pp. 637–641 vol.3, 2000.

[5] M. G. Signorini, A. Fanelli, and G. Magenes, "Monitoring fetal heart rate during pregnancy: Contributions from advanced signal processing and wearable technology," *Comput. Math. Methods Med.*, vol. 2014, 2014.

[6] H. Gonçalves, A. P. Rocha, D. Ayres-de-Campos, and J. Bernardes, "Linear and nonlinear fetal heart rate analysis of normal and acidemic fetuses in the minutes preceding delivery," *Med. Biol. Eng. Comput.*, vol. 44, no. 10, pp. 847–855, 2006.

[7] C. K. Karmakar, A. H. Khandoker, J. Gubbi, and M. Palaniswami, "Complex Correlation Measure: a novel descriptor for Poincaré plot," *Biomed. Eng. Online*, vol. 8, p. 17, Aug. 2009.

[8] S. M. Pincus, “Approximate entropy as a measure of system complexity,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 88, no. 6, pp. 2297–2301, 1991.

[9] J. S. Richman and J. R. Moorman, “Physiological time-series analysis using approximate entropy and sample entropy.,” *Am. J. Physiol. Heart Circ. Physiol.*, vol. 278, no. 6, pp. H2039–H2049, 2000.

[10] W. Chen, Z. Wang, H. Xie, and W. Yu, “Characterization of Surface EMG Signal Based on Fuzzy Entropy,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 15, no. 2. pp. 266–272, 2007.

[11] J. de Haan, J. H. van Bemmel, B. Versteeg, A. F. L. Veth, L. A. M. Stolte, J. Janssens, and T. K. A. B. Eskes, “Quantitative evaluation of fetal heart rate patterns,” *Eur. J. Obstet. Gynecol.*, vol. 1, no. 3, pp. 95–102, 1971.

[12] M. G. Signorini, G. Magenes, S. Cerutti, and D. Arduini, “Linear and nonlinear parameters for the analysisof fetal heart rate signal from cardiotocographic recordings,” *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 3. pp. 365–374, 2003.

[13] G. Georgoulas, C. D. Stylios, and P. P. Groumpos, “Predicting the risk of metabolic acidosis for newborns based on fetal heart rate signal classification using support vector machines,” *IEEE Trans. Biomed. Eng.*, vol. 53, no. 5, pp. 875–884, 2006.

[14] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, no. July 1928, pp. 379–423, 1948.

[15] T. Higuchi, “Approach to an irregular time series on the basis of the fractal theory,” *Phys. D Nonlinear Phenom.*, vol. 31, no. 2, pp. 277–283, 1988.

[16] Z. Ghahramani, “An Introduction To Hidden Markov Models and Bayesian Networks,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 15, no. 01, pp. 9–42, 2001.