# Advanced Artificial Neural Network Classification for Detecting Preterm Births Using EHG Records

Paul Fergus*, Ibrahim Idowu, Abir Hussain, Chelsea Dobbins

School of Computer Science

Liverpool John Moores University

Liverpool L3 3AF, UK

Abstract- Globally, the rate of preterm births are increasing, thus resulting in significant health, development and economic problems. Current methods for the early detection of such births are inadequate. Nevertheless, there has been some evidence that the analysis of uterine electrical signals, collected from the abdominal surface, could provide an independent and easier way to diagnose true labour and detect the onset of preterm delivery. Using advanced machine learning algorithms, in conjunction with Electrohysterography signal processing, numerous studies have focused on detecting true labour several days prior to the event. However, in this paper, the Electrohysterography signals have been used to detect preterm births. This has been achieved using an open dataset, which contains 262 records for women who delivered at term and 38 who delivered prematurely. Several new features from Electromyography studies have been utilized, as well as feature-ranking techniques to determine their discriminative capabilities in detecting term and preterm records. Seven different artificial neural networks were then used to identify these records. The results illustrate that the combination of the Levenberg-Marquardt trained Feed-Forward Neural Network, Radial Basis Function Neural Network and the Random Neural Network classifiers performed the best, with 91% for sensitivity, 84% for specificity, 94% for the area under the curve and 12% for the mean error rate.

Keywords— Electrohysterography (EHG); Preterm Delivery; Term Delivery, Classification, Artificial Neural Networks, Area Under the Curve (AUC), Receiver Operating Curve (ROC) and Feature Extraction

*Corresponding author Tel.: +44(0)151 231 2629, Fax: +44(0)1512074594
Email: p.fergus@ljmu.ac.uk

## 1. Introduction:

The World Health Organisation (WHO) defines preterm birth as the delivery of any baby born alive before 37 weeks of gestation. In other words, births that occur before 259 days of pregnancy are defined as preterm and births that occur between 259 and 294 days, term (WHO, 2012). Preterm births have a significant adverse impact on the new born, including an increased risk of death and other health defects. In particular, infant death rates (less than 24 weeks) are increasing. In 2009, preterm births accounted for approximately 7% of live births, in England and Wales (Bulletin, 2011).

During pregnancy, the monitoring of uterine contractions is vital in order to differentiate between those that are normal and those that may lead to premature birth. The early onset of such contractions can be caused by a number of conditions, including abnormalities in the cervix and uterus, recurrent antepartum haemorrhage and infection (Lucovnik et al., 2011). In the USA, the cost of treatment is reportedly $25.6 billion, whilst in England and Wales, it is estimated to be £2.95 billion, annually (Bulletin, 2011). Consequently, in the last twenty years, a great deal of research has been undertaken to detect and prevent the threat of preterm birth. This has been achieved using different monitoring techniques, which detect uterine contractility. Many approaches have focused on the use of external Tocography and Intrauterine Pressure Catheters. However, they have proven ineffective in the detection of preterm births.

One promising technique, which has gained recognition in monitoring uterine activity, is the use of advanced machine learning algorithms and Electrohysterography (EHG) signal processing. This method records signals from the abdominal surface of pregnant women. These readings are then used to study the electrical activity produced by the uterus. The results are convincing, suggesting that it is an interesting line of enquiry to pursue.

In conjunction with EHG signal processing, the research in Lucovnik et al. (Lucovnik et al., 2011) and Hassan et al. (Hassan, Muszynski, Alexandersson, & Marque, 2013) illustrates that extracting features from EHG signals is key to finding particular spectral information that is specific to term and preterm deliveries. The aim

of this paper is to evaluate the use of selected features in conjunction with several advanced artificial neural network classification algorithms and their ability to distinguish between term and preterm births.

There are several features of the artificial neural networks which make them attractive to medical data classifications. First, artificial neural networks are data driven in that there is no need to make prior assumptions about the model under study. This means that neural network are well suited to problems where their solutions requires some knowledge that is difficult to specify however there enough data or observations. Second, neural network can generalise (Huang, 1998). This means that after the training, they often can produce good results even if the training data contains unseen input patterns. Third, neural networks with the flexible parallel structures can obtain simultaneously the problem solutions (Huang, 2004).

An open dataset has been used, which contains 300 records of pregnant subjects (262 term and 38 preterm). To enable classification, several features have been extracted from the raw EHG signals. These features have not been previously considered in preterm studies. The results indicate that the selected classifiers, in conjunction with the new features, outperform a number of previous approaches.

The remainder of the paper is structured as follows. Section 2 discusses related studies in this field. Section 3 describes the experimental methodology and the selected extracted feature sets, including the design of the experiment. The results have been presented in section 4 and are discussed in Section 5, before the paper is concluded in Section 6.

## 2. Related Studies

Over the past 20 years, an extensive amount of research has focused on the use of pattern recognition techniques to extract features from EHG signals. These include linear and nonlinear methods, in both the time and frequency domains, to improve the results obtained from classification algorithms. The extraction of features often forms part of the data pre-processing stage. In one study, Zardoshti et al. (M. Zardoshti, B. C. Wheeler, K. Badie, 1993), evaluated a number of features commonly used when dealing with EHG signals. These include integrated absolute value, zero crossings and auto-regression coefficient. However, despite their good

discriminant capabilities, a precise frequency threshold for accurate contraction distinction and delivery prediction, over different patients, could not be determined. In our previous work (Paul Fergus et al., 2013), features such as peak frequency, median frequency, root mean square and sample entropy, performed particularly well when discriminating between term and preterm records. Furthermore, several studies have also mentioned very good results, within their reporting, using the same features.

However, it is in the Electromyography (EMG) domain that we find new and interesting works. In one study, Lucovnik et al. (Lucovnik et al., 2011) investigated whether uterine EMG could be used to evaluate propagation velocity (PV). In this study, the electrical signals of the uterus were measured both in labour and non-labour patients who delivered at term and prematurely. The results indicate that, the combination of power spectrum (PS) and PV peak frequency parameters yielded the best predictive results in identifying true preterm labour. However, only one dimension of propagation is considered at a time, which is based on the estimation of time delays between spikes. In comparison, Lange et al. (Lange et al., 2014) estimate the PV of the entire EHG burst that occurs during a contraction. This has been achieved by calculating the bursts corresponding to a full contraction event. The results illustrate that the estimated average propagation velocity is 2.18 (60.68) cm/s. No single preferred direction of propagation was found.

Meanwhile, Alamedine et al. (Alamedine, Khalil, & Marque, 2013) have presented three techniques to identify the most useful features relevant for contraction classification. These include linear features, such as peak frequency, mean frequency and root mean square, and nonlinear features, such as the Lyapunov exponent and sample entropy. In order to choose the most suitable features that represent contractions, feature selection algorithms have also been used. This process involved using a binary particle swarm-optimization (BPSO) algorithm and calculating the Jeffrey Divergence (JD) distance. This is a sequential forward selection (SFS) algorithm. The results show that the BPSO and SFS algorithms could select features with the greatest discriminant capabilities. In this case, out of the six features considered, sample entropy produced the best results.

There is increased interest in detecting term and preterm labour earlier, using non-linear EMG and EHG signals. In one example, Diab et al. (Diab, Hassan, Karlsson, & Marque, 2013) used four non-linear features to distinguish between pregnancy and labour contractions. These features were time reversibility, sample entropy, Lyapunov exponents and delay vector variance. The results show that time reversibility produced the highest classification rate.

In comparison, SooYoung et al. (Sim, Ryou, Kim, Han, & Park, 2014) have used 26 features in their experiment. These include 18 time domain features and 8 frequency domain features. The features have been extracted from 40 signals in the TPEHG database to determine the characteristic differences in uterine muscle activities between term and preterm delivery. The signals are divided into four groups depending on the time of recording (before or after the 26th week of gestation) and the length of gestation (term delivery ≥ 37 weeks and preterm delivery < 37 weeks). The results show significant differences between term and preterm records before 26 weeks when, Frequency Ratio (FR) and Mean Absolute Value Slope1 (MAVSLP1) are used. While other features, such as Willison amplitude (WAMP), Slope Sign Change (SSC), and 3rd Spectral Moments (SM3) show substantial differences between preterm and term delivery data recorded during the later period of gestation.

Yiyao et al (Ye-Lin, Garcia-Casado, Prats-Boluda, Alberola-Rubio, & Perales, 2014) have developed a tool that provides automatic segmentation of EHG recordings, whilst distinguishing between uterine contractions and other artefacts. This has been achieved by using an algorithm that generates the Tocography TOCO signal, derived from the EHG, and detects windows with significant changes in amplitude. In order to develop the classifier, a total of eleven spectral, temporal, and nonlinear features were extracted from the EHG signal windows of 12 women, which were classed by experts as being in the first stages of labour. The combination of characteristics that led to the highest degree of accuracy in detecting artefacts was then determined. Using only seven features, the results produced a precision of 92.2%. This study determined that it is possible to obtain automatic detection of motion artefacts in segmented EHG recordings.

Furthermore, Venugopal et al (Venugopal, Navaneethakrishna, & Ramakrishnan, 2014) have attempted to analyse surface electromyography (sEMG) signals in patients with and without muscle fatigue, using multiple time window (MTW) features. In their experiment, sEMG signals were recorded from the muscles in the biceps brachii of fifty volunteers. Using four window functions (rectangular, Hamming, trapezoidal, and Slepian windows), eleven multiple time window features were acquired. These were selected using a genetic algorithm and information gain based ranking. In addition to this experiment, four different algorithms (naïve Bayes, support vector machines, k-nearest neighbour, and linear discriminant analysis) have also been evaluated to see the impact of the features on each of the classifiers. The results show that, under fatigue, there was a reduction in mean and median frequencies of the signals. The k-nearest neighbour algorithm was the most precise in classifying the features, with a maximum accuracy of 93%.

Meanwhile, Vasak et al. (Vasak et al., 2013) studied whether uterine EMG can identify inefficient contractions. This can lead to first-stage labour and caesarean delivery in term nulliparous women, with the unplanned onset of labour. In this study, EMG was recorded during spontaneous labour in 119 such cases, with singleton term pregnancies in the cephalic position. Electrical activity of the myometrium, during contractions, is characterized by its power density spectrum (PDS). The diagnosis of labour has been made if the patient was in active labour, with no increase in dilation, for at least two hours. The data was analysed to calculate the Intra-class correlation coefficients. This has been achieved by comparing the variance of contraction characteristics, within subjects, to the variance between subjects. The result illustrated that mean peak frequency in women undergoing caesarean delivery, for first-stage labour, was significantly higher (0.55 Hz), than in women delivering vaginally without (0.49 Hz) or with (0.51 Hz) augmentation of labour ($P = .001$ and $P = .01$, respectively). Augmentation of labour increased the mean PDS frequency when comparing contractions before and after the start of augmentation. This increase was only significant in women who eventually delivered vaginally. However, the paper fails to use additional aspects of intra-partum recordings into vitro analysis, testing the hypothesis of a link between an increase in peak frequency and lactic acidosis and impaired in vitro contractility. Furthermore, it also fails to consider other parameter analysis subsets (i.e. sample entropy, root

mean square or wavelet). This could be because, depending on the dataset and parameter analysis equation, the use of different parameter analysis techniques is more challenging in getting meaningful EMG signals. Additionally, if these methods had been applied effectively, it would have led to greater classification results.

## 3. Methodology

This paper uses the TPEHG dataset, which contains the raw EHG signals that are necessary for our study (PhysioNet, 2012). This data has been pre-processed using data segmentation, feature extraction and classification. The study in (Leman H, Marque C, 1999) illustrates how EHG signals can be pre-processed using various frequency related parameters. The study uses several linear and non-linear signal pre-processing techniques, via three different channels, to discern term and preterm deliveries. The pre-processing technique used in (Leman H, Marque C, 1999) passed the EHG signal through a Butterworth filter configured to filter 0.8-4 Hz, 0.3-4 Hz, and 0.3-3Hz frequencies. However, (Maner, 2003) found that uterine electrical activity occurred within 1Hz and that the maternal heart-rate was always higher than 1Hz. Furthermore, 95% of the patients measured had respiration rates of 0.33 Hz or less. Based on these findings, in this paper, the raw TPEHG signals have been passed through the same Butterworth filter to focus on data between 0.34 and 1Hz.

### 3.1 Raw Data Collection

The raw EHG signals, obtained from the Physionet database (PhysioNet, 2012), have been recorded using four bipolar electrodes. These have been adhered to the abdominal surface and spaced at a horizontal and vertical distance between 2.5 and 7cm apart. The total number of records in the EHG dataset is 300 (38 preterm records and 262 term records). Each of the signals have been either recorded early, <26 weeks (at around 23 weeks of gestation) or later, =>26 weeks (at around 31 weeks). Within the dataset, three signals have been obtained simultaneously, ‘per record’. This has been achieved by recording through three different channels.

### 3.2 Feature Extraction

In this paper, several feature extraction techniques have been utilized from (Angkoon Phinyomark, 2009), (Phinyomark, A. Nuidod, P.Phukpattaranont, 2012), (Fele-Zorz, Kavsek, Novak-Antolic, & Jager, 2008) to extract features from the records on channel 3. Table 1, below, describes the features that have been used. In this list, $x_n$ represents the $n^{th}$ sample in the EHG signals in the segment; P represents the power spectrum (calculated using the Fast Discrete Fourier Transform), while N denotes the length of the EHG signal. The main difference between our work and (Angkoon Phinyomark, 2009; Phinyomark, A. Nuidod, P.Phukpattaranont, 2012) is in the analysis of the electrical activity in the uterus, rather than other muscle activity. Given that the uterus is a muscle, this paper investigates whether techniques used to capture EMG activity can also work as well on EHG activity.

### 3.3 Feature Selection

Using the features defined in Table 1, feature vectors have then been generated. The literature reports that peak frequency, median frequency, sample entropy and root mean squares have the most potential to discriminate between term and preterm records. Furthermore, the literature also reports that in EMG studies, the features described in Table 1 are equally as good at discriminating between different muscle activities. However, there is no mention of the uterus in many studies on EMG. To validate these findings, the discriminate capabilities of all the features reported in Table 1 (i.e. feature ranking) have been determined. This has been achieved using several measures, including statistical significance, linear discriminant analysis using independent search (LDAi), linear discriminant analysis using forward search (LDAf), linear discriminant analysis using backward search (LDAb) and gram-schmidt (GS) analysis. Using these measures, the features have been ranked, and the top four uncorrelated features have been selected from the feature space.

Table 1: Feature Extraction Techniques used in EMG

| Equation Name | Mathematic Abbreviation |
|---|---|
| Integrated EMG | $\mathrm{IEMG} = \sum_{n=1}^{N} \lvert(x_n)\rvert$ |
| Mean absolute value of EMG | $\mathrm{MAV} = \frac{1}{N}\sum_{n=0}^{N} \lvert x_n\rvert$ |

| | |
|---|---|
| Simple Square Integral of EMG | $\mathrm{SI} = \sum_{n=0}^{N} \lvert x_n \rvert^2$ |
| Wavelet length of EMG Signal | $\mathrm{WL} = \sum_{n=0}^{N-1} \lvert x_n - x_{n-1} \rvert$ |
| Log Detector of EMG Signal | $\mathrm{LOG} = e^{1/N \sum_{n=1}^{N} \log(\lvert x_n \rvert)}$ |
| Root Mean Square of EMG Signal | $\mathrm{RMS} = \sqrt{1/N \sum_{n=1}^{N} x_n^2}$ |
| Variance of EMG | $\mathrm{VAR} = \frac{1}{N} - 1 \sum_{n=1}^{N} x_n^2$ |
| Difference Absolute Standard Deviation Value of EMG Signal | $\mathrm{DAS} = {}^{1}/_{N-1} \sum_{n=1}^{N-1} (x_{n+1} - x_n)^2$ |
| Maximum Fractal Length of EMG Signal | $\mathrm{MFL} = \log 10(\sqrt{\sum_{n=1}^{N-1} (x_n - x_{n+1})^2})$ |
| Average Amplitude Change of EMG Signal | $\mathrm{AAC} = \frac{1}{N} \sum_{n=1}^{N-1} \lvert x_{n+1} - x_n \rvert$ |
| Peak Frequency of EMG Signal | $\mathrm{fmax} = \arg(\frac{f_s}{N} \max_{i=0}^{N-1} P(i))$ |
| Median Frequency | $\mathrm{fmed} = i_m \frac{f_s}{N}, \quad \sum_{i=0}^{i=i_m} P(i) \doteq \sum_{i=i_m}^{i=N-1} P(i)$ |

These four features have been used in the classification stage to determine which set produced the greatest area under the curve (AUC), sensitivity and specificity values. Table 2 illustrates that the best performing classifier was the Radial Basis Function Neural Network (RBNC), using the Linear Discriminant Analysis Forward Search feature ranking technique. This classifier achieved the best result using the features, sample entropy, waveform length, log detector, and variance.

Table 2: Results for Feature Selection Techniques

| **AUCs for Feature Selection Techniques** | | | | |
|---|---|---|---|---|
| *RBNC* | *RBNC* | *RBNC* | *RBNC* | *RBNC* |
| *p* | *LDAi* | *LDAf* | *LDAb* | *GS* |
| 85% | 87% | 89% | 85% | 87% |

| Sensitivities for Feature Selection Techniques | | | | |
|---|---|---|---|---|
| *RBNC* | *RBNC* | *RBNC* | *RBNC* | *RBNC* |
| *p* | *LDAi* | *LDAf* | *LDAb* | *GS* |
| 79% | 89% | 86% | 81% | 84% |

| Specificities for Feature Selection Techniques | | | | |
|---|---|---|---|---|
| *RBNC* | *RBNC* | *RBNC* | *RBNC* | *RBNC* |
| *p* | *LDAi* | *LDAf* | *LDAb* | *GS* |

| | | | | |
|---|---|---|---|---|
| 74% | 74% | 79% | 74% | 78% |

As a result, this set of features have been used to evaluate the capabilities of the classifiers considered in this paper.

### 3.4 Oversampling of EHG Signals

The TPEHG dataset is unbalanced and contains 262 term and 38 preterm records. This has a significant impact on machine learning algorithms, as classifiers are more prone to detecting the majority class. Therefore, given that there are more term records, the probability of detecting a preterm record is low. To address this issue, the minority class (preterm) is oversampled using the Synthetic Minority Over-Sampling Technique (SMOTE). The technique is effective in solving class skew problems (Richman & Moorman, 2000). Using the 38 preterm records that are already available, SMOTE has been utilized to generate 262 preterm records. The oversampled results have then been compared with the original feature set extracted from the original TPEHG database (262 term and 38 preterm).

### 3.5 Validation Method Used in Experiment

In order to determine the overall accuracy of each of the classifiers several validation techniques have been considered. These include Holdout Cross-validation, K-fold Cross-validation, Sensitivities, Specificities, Receiver Operating Curve (ROC) and Area Under the Curve (AUC).

### 3.6 Classifiers

This study evaluates the use of seven advanced artificial neural network classifiers. This includes the back-propagation trained feed-forward neural network classifier (BPXNC), levenberg-marquardt trained feed-forward neural network classifier (LMNC), the perceptron linear classifier (PERLC), radial basis function neural network classifier (RBNC), random neural network classifier (RNNC), the Voted Perceptron classifier (VPC) and the Discriminative Restricted Boltzmann Machine classifier (DRBMC) (37steps, 2013).

In the BPXNC, the network is trained to map a set of input data by iterative adjustment of the weights. The information from inputs is fed forward through the network to optimize the weights between neurons. Moreover, the optimization of the weights is made by backward propagation of the error during the training or learning stage. The BPXNC then reads the input and output values in the training dataset and changes the value of the weighted links to reduce the differentiation between the predicted and observed values. The error in prediction is reduced across several training cycles (epoch 50) until the network reaches the best level of classification accuracy, while avoiding overfitting (Ghaffari et al., 2006).

The Levenberg-Marquardt trained feed-forward neural network classifier (LMNC) is similar to the BPXNC, in terms of functionality. However, it is much more memory intensive. Furthermore, during the training stage, training is stopped when the performance on an artificially generated tuning set of 1000 samples per class has been reached (based on k-nearest neighbour interpolation) and thereafter does not improve (37steps, 2013).

Linear perceptron linear classifiers (PERLC) are the simplest type of neural network classifier and are trained with a supervised training algorithm. This classifier assumes that the true classes of the training data are available and incorporated in the training process. The input weights in this classifier can be adjusted iteratively by the training algorithm so as to produce the correct class mapping for the output. However, the problem with this classifier is that it does not have a hidden layer therefore this leads to bias in result accuracy.

The radial basis function neural network classifier (RBNC) is mostly used in complicated pattern recognition and classification problems, such as biomedical datasets that are nonlinear (Huang, 1999). The classifier has one hidden layer with unit radial basis units. The mapping properties of the RBCN can be modified through the weights in the output layer.

The Random neural net classifier (RNNC) is a feed-forward neural network with one hidden layer consisting of N sigmoid neurons. The input layer rescales the input features to unit variance; the hidden layer has normally distributed weights and biases with zero mean and standard deviation (37steps, 2013).

The voted perceptron classifier is an improved version of perceptron networks which was proposed by (Freund and Schapire, 1999). The algorithm takes advantages of the data that are linearly separable with a large margin. Similar the support vector machine, the network can be used with the kernel function.

Discriminative Restricted Boltzmann Machine classifier (DRBMC) is a powerful classifier based on latent variables which are usually binary numbers for the modelling of input distributions (Larochelle, Bengio, 2008). In this case, the variables in the visible layer are separated into two parts. The first represents the input data and the second represents the label of input.

## 4. Results

This section presents the classification results for term and preterm delivery records. This has been achieved using the extracted feature set from the 0.34-1 Hz filter on Channel 3. Using the 80% holdout technique and k-fold cross-validation, the initial validation results are presented. This provides a baseline for comparison against all subsequent evaluations that have been performed, using the oversampled dataset, clinical data and the combination of classifiers.

### 4.1 Original Results for 0.34-1 Hz Filter on Channel 3

The performance of each classifier has been evaluated using the mean sensitivity, specificity, errors, standard deviation, and AUC values. Each experiment has been repeated 30 times, with randomly selected training and test sets for each run.

#### Classifier Performance

The first evaluation uses the original TPEHG dataset, which contains 38 preterm and 262 term observations. Table 3, illustrates the mean averages obtained over 30 simulations for the sensitivity, specificity, and AUC values. As it can be noticed, the sensitivities (i.e. the ability to classify a preterm record), in this initial test, are low for all classifiers. This is expected since the dataset is unbalanced in favour of term observations, thus there are a limited number of preterm records from which the classifiers can learn. Consequently, specificities are much higher than sensitivities.

Table 3: Original TPEHG Signal (262 Term And 38 Preterm)

| Classifiers | Sensitivity | Specificity | AUC |
|---|---|---|---|
| **BPXNC** | 0.0000 | 0.9987 | 54% |
| **LMNC** | 0.0667 | 0.9519 | 58% |
| **PERLC** | 0.1619 | 0.8647 | 57% |
| **RBNC** | 0.1286 | 0.9622 | 56% |
| **RNNC** | 0.0667 | 0.9474 | 56% |
| **VPC** | 0.0000 | 1.0000 | 50% |
| **DRBMC** | 0.0000 | 0.9981 | 58% |

Table 4, illustrates the results obtained from k-fold cross-validation. This has been used to determine whether the results from the holdout method can be improved.

Table 4: Original TPEHG signal (262 Term and 38 Preterm) cross-validation

| 80% Holdout: 30 Repetitions | | | Cross Val, 5 Folds, 1 Repetition |
|---|---|---|---|
| **Classifiers** | **Mean Err** | **Standard Deviation** | **Mean Err** |
| **BPXNC** | 0.1278 | 0.0043 | 0.1333 |
| **LMNC** | 0.1602 | 0.0331 | 0.1767 |
| **PERLC** | 0.2243 | 0.1186 | 0.2400 |
| **RBNC** | 0.1434 | 0.0342 | 0.1333 |
| **RNNC** | 0.1641 | 0.0363 | 0.1567 |
| **VPC** | 0.1267 | 0.0000 | 0.1267 |
| **DRBMC** | 0.1283 | 0.0068 | 0.1267 |

The k-fold cross-validation results, using five folds and one repetition illustrate that k-fold cross-validation has improved the error rates, for some of the classifiers. However, these are negligible. Furthermore, the lowest error rates could not be improved below the minimum error rate expected, which is 12.67%.

### 4.2 Results for 0.34-1 Hz TPEHG filter on Channel 3 – Oversampled using SMOTE

In order to improve the results, the preterm observations have been oversampled using SMOTE technique. This algorithm balances the dataset by oversampling the

minority class (38 preterm records) to 262. A new dataset is generated that contains an even split between term and preterm records.

Classifier Performance

Table 5 illustrates that the sensitivities, for all of the algorithms, have significantly improved, while specificities have decreased. In addition, the AUC results also show a significant improvement in accuracy for all of the classifiers. In particular, the RBNC classifier has dramatically improved with an accuracy of 90%.

Table 5: SMOTE TPEHG signal (262 Term and 262 Preterm)

| Classifiers | Sensitivity | Specificity | AUC |
|---|---|---|---|
| **BPXNC** | 79% | 58% | 72% |
| **LMNC** | 82% | 69% | 82% |
| **PERLC** | 46% | 67% | 63% |
| **RBNC** | 85% | 80% | 90% |
| **RNNC** | 86% | 72% | 83% |
| **VPC** | 98% | 2% | 50% |
| **DRBMC** | 59% | 55% | 56% |

Table 6: SMOTE TPEHG signal (Term and Preterm) cross-validation

| | 80% Holdout: 30 Repetitions | | Cross Val, 5 Folds, 1 Repetition |
|---|---|---|---|
| **Classifiers** | **Mean Err** | **Standard Deviation** | **Mean Err** |
| **BPXNC** | 0.3144 | 0.0591 | 0.2977 |
| **LMNC** | 0.2455 | 0.0489 | 0.2195 |
| **PERLC** | 0.4321 | 0.0624 | 0.4656 |
| **RBNC** | 0.1734 | 0.0424 | 0.1622 |
| **RNNC** | 0.2106 | 0.0451 | 0.2023 |
| **VPC** | 0.4984 | 0.0088 | 0.5000 |
| **DRBMC** | 0.4295 | 0.0376 | 0.4198 |

Table 6 illustrates the resulting mean error rates of the oversampled dataset. As it can be seen, the mean error rates, produced by all of the classifiers, are lower than the cross-validation mean errors and the expected error rate, which is 262/524, i.e. 50%.

### 4.3 Results for 0.34-1 Hz TPEHG filter on Channel 3 combined with Clinical Data

Clinical data for each of the women in the dataset were made available in December 2012. These include the age of the women, parity (the number of previous births), abortions, weight, hypertension, diabetes, placental position, first and second trimester bleeding, funnelling and smoking. Once the clinical data has been added to the original dataset, several observations were removed because of missing clinical data. This resulted in a new dataset containing 17 preterm records and 152 term records. Again, in order to balance the dataset, the preterm records have been oversampled using SMOTE to produce 153 preterm and 152 term samples. A new dataset is created that combines the real and synthetic observations (305 observations altogether). Using this dataset, the experiment is again performed using 30 iterations.

#### Classifier Performance

Table 7 illustrates that the results have improved slightly to those presented in Table 5. Several of the classifiers have now produced higher values for the AUC and both the sensitivities and specificities. This is despite having to reduce the size of the dataset to account for missing values in the clinical data (in the case of preterm 22 observations had to be removed; and in the case of term 110 observations had to be removed).

Table 7: SMOTE TPEHG signal (152 Term and 153 Preterm) with Clinical Data

| Classifiers | Sensitivity | Specificity | AUC |
|---|---|---|---|
| **BPXNC** | 64% | 64% | 68% |
| **LMNC** | 85% | 76% | 85% |
| **PERLC** | 53% | 61% | 64% |
| **RBNC** | 87% | 81% | 91% |
| **RNNC** | 87% | 71% | 84% |
| **VPC** | 100% | 0% | 50% |
| **DRBMC** | 56% | 55% | 52% |

Table 8 illustrates the resulting mean error rates of the dataset containing the clinical data. As it can be seen, the mean error rates produced by several of the

classifiers, are much lower than the expected error rate, which is 153/304, i.e. 50%, and are comparable with the cross-validation mean errors.

Table 8: SMOTE TPEHG signal (Term and Preterm) with Clinical Data cross-validation

| | 80% Holdout: 30 Repetitions | | Cross Val, 5 Folds, 1 Repetition |
|---|---|---|---|
| **Classifiers** | **Mean Err** | **Standard Deviation** | **Mean Err** |
| **BPXNC** | 0.3594 | 0.0839 | 0.3508 |
| **LMNC** | 0.1932 | 0.0710 | 0.1803 |
| **PERLC** | 0.4329 | 0.0674 | 0.3639 |
| **RBNC** | 0.1643 | 0.0365 | 0.1377 |
| **RNNC** | 0.2097 | 0.0460 | 0.1934 |
| **VPC** | 0.4984 | 0.0000 | 0.4656 |
| **DRBMC** | 0.4444 | 0.0561 | 0.4525 |

## 4.4 Results for 0.34-1 Hz TPEHG filter on Channel 3 with Additional Features and Clinical Data

Building on our previous work (Fergus et al., 2013), this experiment combines features from that work that produced good results. These additional features are root mean squares, peak frequency and median frequency.

### Classifier Performance

Table 9 illustrates that the results have improved on those presented in Table 7, indicating that the additional features provide better separation between the two classes.

Table 10 illustrates the resulting mean error rates of the dataset containing the clinical data. As it can be seen, the mean error rates, produced by several of the classifiers, are much lower than the expected error rate, which is 50%, and comparable with the cross-validation mean errors.

Table 9: SMOTE TPEHG signal (152 Term and 153 Preterm) with Additional Features and Clinical Data

| Classifiers | Sensitivity | Specificity | AUC |
|---|---|---|---|
| **BPXNC** | 67% | 67% | 70% |
| **LMNC** | 95% | 81% | 88% |
| **PERLC** | 56% | 62% | 65% |
| **RBNC** | 70% | 95% | 94% |
| **RNNC** | 88% | 72% | 87% |
| **VPC** | 100% | 0% | 50% |
| **DRBMC** | 61% | 51% | 51% |

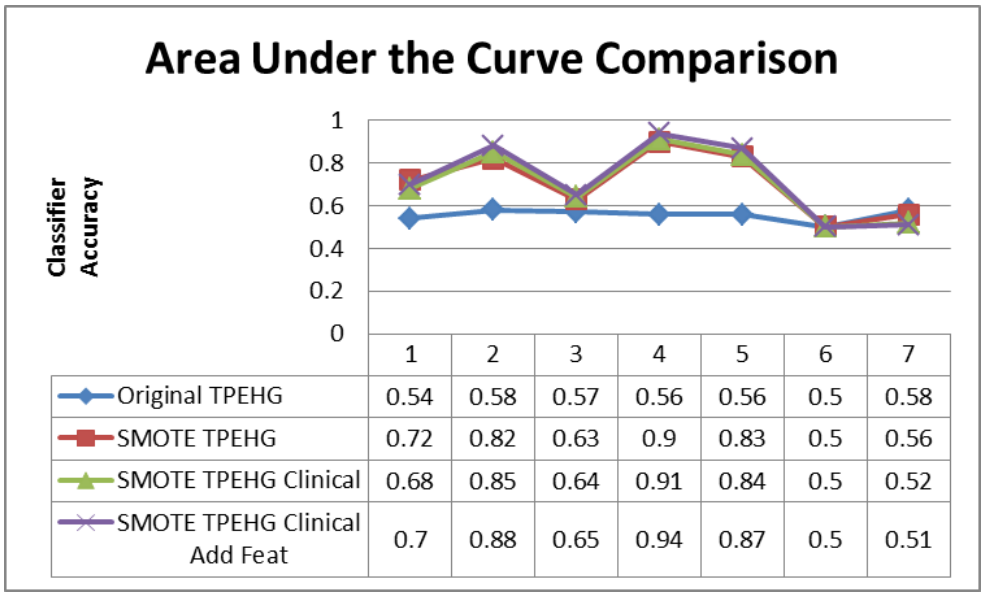
Table 10: SMOTE TPEHG signal (Term and Preterm) with Additional Features and Clinical Data cross-validation

| | 80% Holdout: 30 Repetitions | | Cross Val, 5 Folds, 1 Repetition |
|---|---|---|---|
| **Classifiers** | **Mean Err** | **Standard Deviation** | **Mean Err** |
| **BPXNC** | 0.3317 | 0.0838 | 0.3508 |
| **LMNC** | 0.1220 | 0.0560 | 0.1803 |
| **PERLC** | 0.4118 | 0.0542 | 0.3639 |
| **RBNC** | 0.1749 | 0.0406 | 0.1377 |
| **RNNC** | 0.1992 | 0.0451 | 0.1934 |
| **VPC** | 0.4984 | 0.0000 | 0.4656 |
| **DRBMC** | 0.4421 | 0.0527 | 0.4525 |

### 4.5 Classifier Performance Comparison

The results from the previously run experiments have now been compared in Figures 1, 2 and 3. As it can be seen in Figure 1, all of the classifiers have performed consistently under the four different strategies taken. However, the original unbalanced TPEHG dataset does provide the poorest results. This is due to the disparity between term and preterm observations. Interestingly, the linear and voted perceptron classifiers do not provide sufficient models for prediction in any of the strategies used. This is a similar case for the Discriminative Restricted Boltzmann Machine classifier. The simulation results indicate that using the SMOTE oversampling technique, with clinical data and added features, provides the best

AUC using the Radial basis Neural Network classifier. This is followed closely by the Levenberg-Marquardt trained Feed-Forward Neural Network classifier and the Random Neural Network classifier.




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Original TPEHG | 0.54 | 0.58 | 0.57 | 0.56 | 0.56 | 0.5 | 0.58 |
| SMOTE TPEHG | 0.72 | 0.82 | 0.63 | 0.9 | 0.83 | 0.5 | 0.56 |
| SMOTE TPEHG Clinical | 0.68 | 0.85 | 0.64 | 0.91 | 0.84 | 0.5 | 0.52 |
| SMOTE TPEHG Clinical Add Feat | 0.7 | 0.88 | 0.65 | 0.94 | 0.87 | 0.5 | 0.51 |

**Fig 1:** Comparison of AUC values using the Four Strategies. Numbers 1 to 7 represent BPXNC, LMNC, PERLC, RBNC, RNNC, VPC and DRBMC classifiers respectively.

Figure 2 presents the sensitivities and hence the classifiers ability to predict preterm observations. The focus of the paper has been to improve sensitivity rates, as it is more important to predict preterm delivery, as opposed to miss-classifying a term pregnancy. As expected, the sensitivities are low using the original data. This is solely due to the majority of observations being term and only a small number of observations being preterm. The highest sensitivity readings have resulted from strategies 2, 3 and 4, using the Levenberg-Marquardt trained Feed-Forward Neural Network classifier and Trainable Radial Basis Neural Network classifier. This is consistent with the AUC values that have been depicted in Figure 1. Interestingly, the sensitivities are high for the Voted Perceptron classifier, yet the findings are inconsistent with the very low AUC values for this classifier in Figure 1.

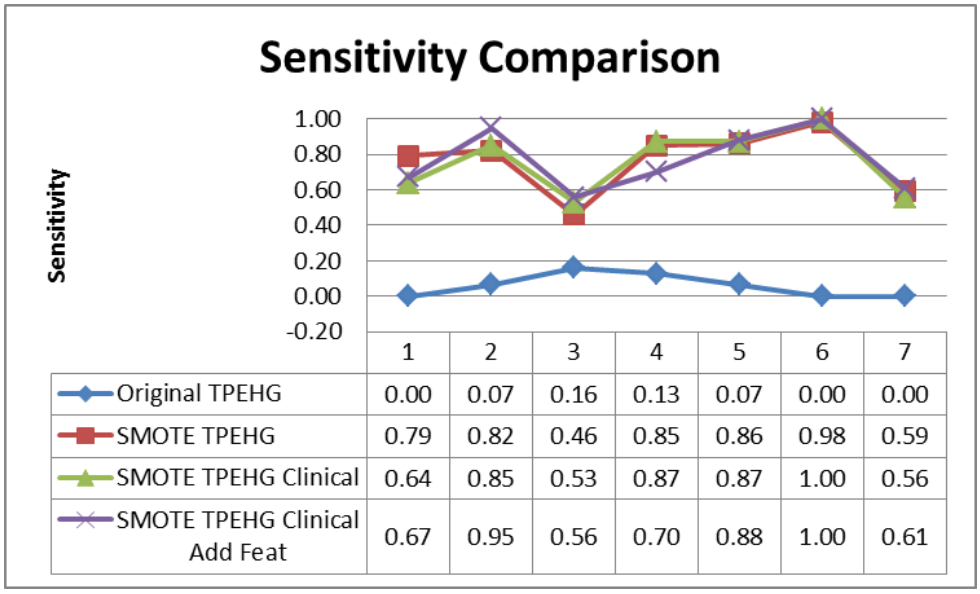



**Fig 2:** Comparison of Sensitivity values using the Four Strategies. Numbers 1 to 7 represent BPXNC, LMNC, PERLC, RBNC, RNNC, VPC and DRBMC classifiers, respectively

Lastly, Figure 3 illustrates the specificity results for each of the strategies that have been used. As expected, the specificity values for all classifiers, using strategy one, are high. Again, this is due to the unbalanced dataset (i.e. 262 out of the 300 observations were term). For the Levenberg-Marquardt trained Feed-Forward Neural Network and the Radial Basis Neural Network classifiers, the values are consistent with the previous figures. Interestingly, using strategy three and four, it is the Levenberg-Marquardt trained Feed-Forward Neural Network classifier that performed better at predicting preterm, whilst the Radial Basis Neural Network classifier is better at predicting term.

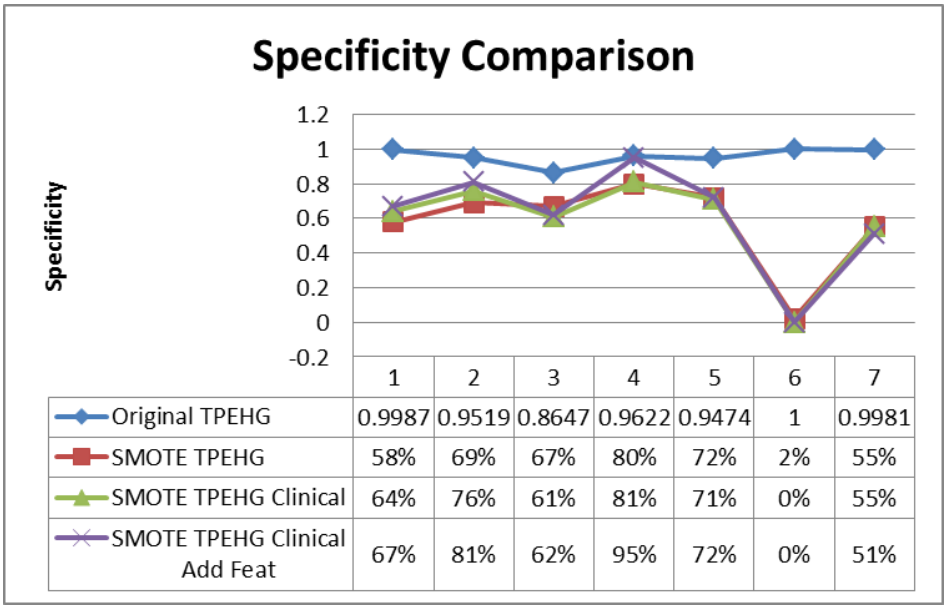



**Fig 3:** Comparison of Specificity values using the Four Strategies. Numbers 1 to 7 represent BPXNC, LMNC, PERLC, RBNC, RNNC, VPC and DRBMC classifiers, respectively

### 4.6 Combining Classifiers

In this paper, another set of experiments has been conducted to determine whether the results can be further improved. This involved combining the best classifiers that produced consistent AUC, Sensitivity and Specificity values across all of the strategies. The classifiers that fulfil this criterion are the Levenberg-Marquardt trained Feed-Forward Neural Network, Radial Basis Neural Network classifier and the Random Neural Network classifier.

#### Classifier Performance

Table 11, illustrates that the results can be improved further on those presented in Table 9, with several of the classifiers producing higher values for the AUC and both the sensitivities and specificities. This suggests that by combining the predictive capabilities from each classifier, better results can be obtained.

Table 11: Combined Classifiers

| Classifiers | Sensitivity | Specificity | AUC |
|---|---|---|---|
| **LMNC, RBNC, and RNNC Combined** | 91% | 84% | 94% |

Table 12 illustrates that there is a 12% error, which is slightly high, but much lower than the *expected error rate*. The cross validation results demonstrate that the 80% holdout technique produces the better results.

Table 12: SMOTE TPEHG signal (Term and Preterm) with Additional Features and Clinical Data cross-validation

| | 80% Holdout: 30 Repetitions | | Cross Val, 5 Folds, 1 Repetition |
|---|---|---|---|
| **Classifiers** | **Mean Err** | **Standard Deviation** | **Mean Err** |
| **LMNC, RBNC and RNNC Combined** | 0.1254 | 0.0521 | 0.1623 |

The results illustrate that using machine learning techniques are encouraging. Within a wider context, this approach might be able to utilise real-life pregnancy data to predict, with high confidence, whether an expectant mother is likely to have a premature birth or proceed to full term.

## 5. Discussion

Most of the uterine EHG signal studies concentrate on predicting true labour, which is based on the last stage of the pregnancy duration. This paper has studied the uterine EHG signals of women in order to classify the preterm and term deliveries from the early stages of the pregnancy. It has been suggested that ANN is a better solution for nonlinear medical decision support systems than traditional statistical techniques (Li et al. 2000). Therefore, this experiment is based on applying seven different types of neural networks for the classification of term/preterm data.

The initial classification with the data set in its original form achieved very low sensitivity, below 20%, while the specificity is higher. This means that the classifiers were classifying most of the cases into the majority class, which are term subjects. The main reason for the ineffective classification was the unequal amount of term records to preterm records. Therefore, in these experiments, the oversample SMOTE method has significantly improved the sensitivity and specificity rates for most of the ANN classifiers.

The first publication of the TPEHG data set was in 2010. However, additional clinical data became freely available in 2012. The additional features from the TPEHG database with the clinical data were considered in our experiments when analysing the data set. The experiment results demonstrate that the general performance of most ANN classifiers is significantly improved further by comprising the information from the clinical data set.

By combining additional features with the clinical data, our simulation results showed further improvements in terms of the average sensitivity, specificity and area under the curve. In this case, the results show that the Levenberg-Marquardt trained Feed-Forward Neural Network classifier performs better at predicting preterm records while the Radial Basis Neural Network classifier is better at predicting term. This is clearly indicating that using single classifier for the prediction of term/preterm real data may not generate good results, while combining a number of classifiers can generate more reliable classification.

## 6. Conclusion

The development of medical information systems has played an important role in the biomedical domain. This has led to the extensive use of Artificial Intelligence (AI) techniques for extracting biological patterns in data. Furthermore, data pre-processing and validating techniques have also been used extensively to analyze such datasets for classification problems. In this paper, seven classifiers have been used to classify term and preterm records from the TPEHG dataset, filtered between 0.34 and 1 Hz. The results demonstrate that the best performing classifier was the RBNC with 85% sensitivity, 80% specificity, 90% AUC and a 17% mean error rate. These results are encouraging and suggest that the approach posited in this paper is a line of enquiry worth pursuing.

Perhaps one negative aspect of the work is the need to utilize oversampling to increase the number of preterm samples. A better way would have been to balance the dataset using actual recordings obtained from pregnant women who delivered prematurely. This will be the focus of future research, alongside a more extensive investigation into different machine learning algorithms and techniques.

## References


37steps. (2013). *Pattern Recognition Tools*. *Version 5*.

Alamedine, D., Khalil, M., & Marque, C. (2013). Comparison of different EHG feature selection methods for the detection of preterm labor. *Computational and Mathematical Methods in Medicine*, *2013*, 485684. doi:10.1155/2013/485684

Angkoon Phinyomark, C. L. and P. P. (2009). A Novel Feature Extraction for Robust EMG Pattern Recognition. *Journal of Computing*, *1*(1), 71–79.

Bulletin, S. (2011). *Statistical Bulletin Gestation-specific Infant Mortality in England and Wales*. *National Office for Statistics*.

Diab, A., Hassan, M., Karlsson, B., & Marque, C. (2013). Effect of decimation on the classification rate of nonlinear analysis methods applied to uterine EMG signals. *Utc.fr*, 12–14.

Fele-Zorz, G., Kavsek, G., Novak-Antolic, Z., & Jager, F. (2008). A comparison of various linear and non-linear signal processing techniques to separate uterine EMG records of term and pre-term delivery groups. *Medical & Biological Engineering & Computing*, *46*(9), 911–22. doi:10.1007/s11517-008-0350-y

Freund, Y., Schapire, R. E., (1999). Large margin classification using the perceptron algorithm. Machine Learning, 37 (3), 277-296.

Fergus, P., Cheung, P., Hussain, A., Al-Jumeily, D., Dobbins, C., & Iram, S. (2013). Prediction of Preterm Deliveries from EHG Signals Using Machine Learning. *PloS One*, *8*(10), e77154. doi:10.1371/journal.pone.0077154

Ghaffari, A., Abdollahi, H., Khoshayand, M. R., Bozchalooi, I. S., Dadgar, A., & Rafiee-Tehrani, M. (2006). Performance comparison of neural network training algorithms in modeling of bimodal drug delivery. *International Journal of Pharmaceutics*, *327*(1-2), 126–38. doi:10.1016/j.ijpharm.2006.07.056

Hassan, M., Muszynski, C., Alexandersson, A., & Marque, C. (2013). Nonlinear Correlation Analysis of External Uterine Electromyography. *IEEE Transactions on BioMedical Engineering*, *60*(4), 1160–1166.

Huang, D. S. (1998). The local minima free condition of feedforward neural networks for outer-supervised learning, *IEEE Trans on Systems, Man and Cybernetics*, vol.28B, no.3, 477-480.

Huang, D. S. (2004). A constructive approach for finding arbitrary roots of polynomials by neural networks, *IEEE Transactions on Neural Networks*, vol.15, no.2, 477-491.

Huang, D.S. (1999). Radial basis probabilistic neural networks: Model and application, International Journal of Pattern Recognition and Artificial Intelligence, 13(7), 1083-1101.

Lange, L., Vaeggemose, A., Kidmose, P., Mikkelsen, E., Uldbjerg, N., & Johansen, P. (2014). Velocity and directionality of the electrohysterographic signal propagation. *PloS One*, *9*(1), e86775. doi:10.1371/journal.pone.0086775

Larochelle, H. , Bengio, Y. (2008). Classification using discriminative restricted boltzmann machines, *Proceeding 25th Internaional Conerence. Machine Learing*, 536–543.

Leman H, Marque C, G. J. (1999). Use of the electrohyster- ogram signal for characterization of contractions during pregnancy. *IEEE Trans Biomed Eng*, *46*(10), 1222–1229.

Li, Y.C., Liu, L., Chiu, W.T., Jian, W. S.,(2000). Neural network modeling for surgical decisions on traumatic brain injury patients. I*nternational Journal of Medical Informatics*, 57(1), 1–9.

Lucovnik, M., Maner, W. L., Chambliss, L. R., Blumrick, R., Balducci, J., Novak-Antolic, Z., & Garfield, R. E. (2011). Noninvasive uterine electromyography for prediction of preterm delivery. *American Journal of Obstetrics and Gynecology*, *204*(3), 228.e1–10. doi:10.1016/j.ajog.2010.09.024

M. Zardoshti, B. C. Wheeler, K. Badie, R. H. (1993). Evaluation of EMG Features for Movement Control of Prostheses. In *proceedings.15th International. conference. IEEE EMBS* (pp. 1141–1142).

Maner, W. (2003). Predicting term and preterm delivery with transabdominal uterine electromyography. *Obstetrics & Gynecology*, *101*(6), 1254–1260. doi:10.1016/S0029-7844(03)00341-7

Phinyomark, A. Nuidod, P.Phukpattaranont, C. L. (2012). Feature Extraction and Reduction of Wavelet Transform Coefficients for EMG Pattern Classification. *Electronics & Electrical Engineering*, *6*(6).

PhysioNet. (2012). *The Term -Preterm EHG Database (TPEHG- DB)*. *physionet.org*.

Richman, J., & Moorman, J. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology- …*, *278*(6), H2039–49.

Sim, S., Ryou, H., Kim, H., Han, J., & Park, K. (2014). Evaluation of Electrohysterogram Feature Extraction to Classify the Preterm and Term Delivery Groups. In *Find out how to access preview-only content The 15th International Conference on Biomedical Engineering IFMBE Proceedings* (pp. 675–678). doi:10.1007/978-3-319-02913-9_172

Vasak, B., Graatsma, E. M., Hekman-Drost, E., Eijkemans, M. J., van Leeuwen, J. H. S., Visser, G. H., & Jacod, B. C. (2013). Uterine electromyography for identification of first-stage labor arrest in term nulliparous women with

spontaneous onset of labor. *American Journal of Obstetrics and Gynecology*, *209*(3), 232.e1–8. doi:10.1016/j.ajog.2013.05.056

Venugopal, G., Navaneethakrishna, M., & Ramakrishnan, S. (2014). Extraction and analysis of multiple time window features associated with muscle fatigue conditions using sEMG signals. *Expert Systems with Applications*, *41*(6), 2652–2659. doi:10.1016/j.eswa.2013.11.009

WHO. (2012). *Born too soon: The Global Action Report on Preterm Birth* (p. 126). Retrieved from http://www.who.int/maternal_child_adolescent/documents/born_too_soon/en/index.html

Ye-Lin, Y., Garcia-Casado, J., Prats-Boluda, G., Alberola-Rubio, J., & Perales, A. (2014). Automatic Identification of Motion Artifacts in EHG Recording for Robust Analysis of Uterine Contractions. *Computational and Mathematical Methods in Medicine*, *2014*, 470786. doi:10.1155/2014/470786