A Machine-Learning Approach to the Detection of Fetal Hypoxia during Labor and Delivery

Philip A. Warrick, Emily F. Hamilton, Robert E. Kearney, Doina Precup

■ *Labor monitoring is crucial in modern health care, as it can be used to detect (and help avoid) significant problems with the fetus. In this article we focus on detecting hypoxia (or oxygen deprivation), a very serious condition that can arise from different pathologies and can lead to lifelong disability and death. We present a novel approach to hypoxia detection based on recordings of the uterine pressure and fetal heart rate, which are obtained using standard labor monitoring devices. The key idea is to learn models of the fetal response to signals from its environment. Then, we use the parameters of these models as attributes in a binary classification problem. A running count of pathological classifications over several time periods is taken to provide the current label for the fetus. We use a unique database of real clinical recordings, both from normal and pathological cases. Our approach classifies correctly more than half the pathological cases, 1.5 hours before delivery. These are cases that were missed by clinicians; early detection of this type would have allowed the physician to perform a Cesarean section, possibly avoiding the negative outcome.*

The lifelong disability that can result from oxygen deprivation during childbirth is rare but devastating for families, clinicians, and the health-care system. Between 1 and 7 in 1000 fetuses experience oxygen deprivation during labor that is severe enough to cause fetal death or brain injury (Saphier et al. 1998); the range of this estimate reflects considerable regional variation and some clinical debate on the definition of brain injury. The main source of information used by clinicians to assess the fetal state during labor is cardiotocography (CTG), which measures maternal uterine pressure (UP) and fetal heart rate (FHR). These signal are routinely recorded during labor, using monitors of the type presented in figure 1.

Clinicians look at these signals and use visual pattern recognition and their prior experience to decide whether the fetus is in distress and to pick an appropriate course of action (such as performing a Cesarean section). However, there is great variability among physicians in terms of how they perform this task (Parer et al. 2006). Furthermore, because signicant hypoxia is rare, false alarms are common, leading physicians to disregard truly abnormal signals. Indeed, approximately 50 percent of birth-related brain injuries are deemed preventable, with incorrect CTG interpretation leading the list of causes (Draper et al. 2002; Freeman, Garite, and Nageotte 2003). The social costs of such errors are massive: intrapartum care generates the most frequent malpractice claims and the greatest liability costs

Figure 1. A Standard Bedside Labor Monitor.

The electrodes are attached to the mother's abdomen, and they record the uterine pressure and fetal heart rate (seen on the paper above).

of all medical specialties (Saphier et al. 1998). Thus, there is great motivation to find better methods to discriminate between healthy and hypoxic conditions.

In this article, we summarize our recent work on a novel approach to this problem, which relies heavily on machine-learning methods; a more detailed account of the methods is presented in two biomedical journal publications (Warrick et al. 2009; 2010) as well as in a Ph.D. dissertation (Warrick 2010). Although the system we present was designed specifically for labor monitoring, we believe that the general steps we took can inform other AI medical monitoring systems, as well as, more generally, applications for time-series prediction and analysis.

We built an automated detector of fetal distress by using data from normal and pathological cases. We had access to a unique database, which contains labor monitoring data from an unusually large number of births; a significant number of the cases are pathological examples (well above the natural frequency of occurrence of such problems). All the data has been collected under clinical conditions; as a result, it is very noisy. To handle this problem, we modeled the fetal heart rate signal through several components. The parameters of these models, which have been learned from data, are then used to build a classifier for a given time period. Because the state of the fetus can change during labor, classification is performed repeatedly on data segments of limited duration. A majority vote of recent labels determines if and when a fetus is considered pathological.

The article is organized as follows. First, we give some background on the problem and type of data used. Then, we describe our general approach. We present empirical results and a discussion of the main findings, as well as the next steps towards clinical deployment.

Background

Clinicians' interpretation of intrapartum CTG signals relies on the temporary decreases in FHR (FHR

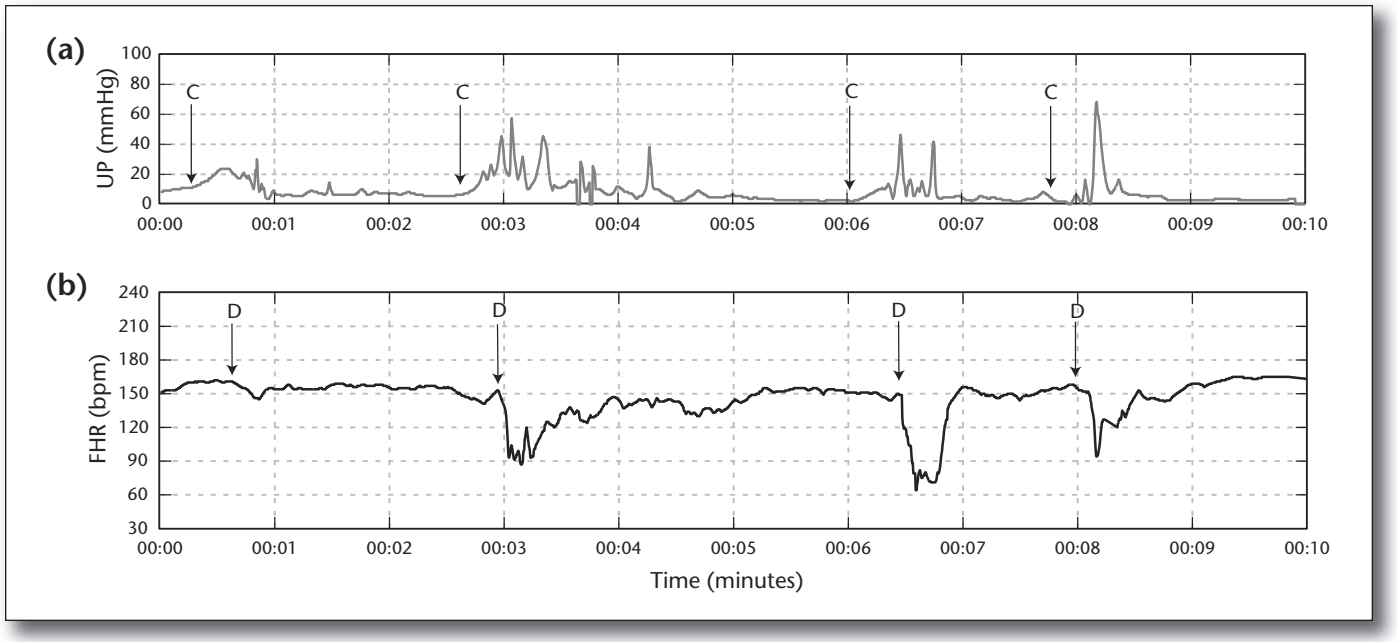

Figure 2. CTG Signal over 10 Minutes, Including Four Contraction-Deceleration Pairs.

Top: UP signal with contraction onsets (C) indicated. *Bottom:* FHR signal with deceleration onsets (D) indicated.

decelerations) in response to uterine contractions. FHR decelerations are due mainly to two contraction-induced events: (1) umbilical-cord compression and (2) a decrease in oxygen delivery through an impaired utero-placental unit. There is general consensus that deceleration depth, frequency, and timing with respect to contractions are indicators of both the insult and the ability of the fetus to withstand it. Figure 2 shows an example CTG during 10 minutes. Contractions and decelerations are marked by an expert.

There have been numerous studies in the literature that describe fetal-state assessment based on computerized interpretation of the CTG signal, for example, the papers by Georgoulas, Stylios, and Groumpos (2006a; 2006b) and Ozyilmaz and Yildirim (2004). By far the majority of these have been based on a paradigm of detection and estimation of attributes selected to mirror the obstetrician's visual interpretation of the UP and FHR graphs, or to reflect assumed physiological events. For example, one can attempt to detect the start and end of a contraction, the start and end of the following deceleration, the depth of the deceleration, and more (Signorini et al. 2003). Georgoulas et al. (2006a, 2006b) use principal component analysis and support vector machines on top of features computed from the heart signal in order to provide a classification for the fetus. Ozyilmaz and Yildirim (2004) use neural networks and radial basis function with features of the heart signal and the gestational age. In similar work in the past (Warrick, Hamilton, and Macieszczak 2005) we have also used combinations of neural networks and feature extraction.

Unfortunately, several problems hamper the use of such features. First, the UP and FHR signals are very noisy, especially when collected under clinical conditions, as is the case for our data. Because the sensors are attached to the maternal abdomen, there is often a problem of sensor contact or missing data when the mother wishes to be more mobile and is temporarily detached from monitoring. These sensor disturbances result in frequent artifacts, where the signal drops to a lower value. The FHR can also include interference from the maternal heart rate, causing the signal to drop to a lower value.

Other problems arise from the fact that detecting events like the start of a deceleration is very hard to do automatically. The response of the fetus is not always the same: while most of the time, a contraction is followed by a deceleration, sometimes it may actually be followed by an acceleration. Furthermore, a missed detection can throw off the timing information for future contractions and skew all subsequent results.

Finally (and perhaps most importantly) looking at features of the FHR in isolation does not give information on how the fetus is reacting to the labor. Pathology is often indicated by the response of the fetus to contractions; but the relationship between the UP and FHR is not captured explicitly in the FHR features.

Because of the problems attributed to feature detection, we decided to build data-driven models of the CTG. Our models are structured based on clinical knowledge, capturing information about the main physiological determinants of the fetal heart rate. However, we do not attempt to mimic a doctor's visual interpretation of the CTG, since this is difficult for an automated system to match. Instead, we focus on the crucial interaction between the uterine pressure, as an input signal, and the fetal heart rate, as an output signal. Doctors take this relationship into account in their visual assessment of contraction-deceleration timing. We use the recorded data to fit the models, without trying to impose prior knowledge of what the relationship of the two signals should be. The parameters of the models are learned from data. There are two potential advantages to this approach. First, as we already stated, we avoid the difficulty of trying to mimic visual interpretation. Second, it is much easier to detect changes in the state of the fetus as labor progresses, rather than trying to refer to some "golden standard" as is often done in the feature-based approach. This can allow detection to be more attuned to the individual characteristics of each labor. Once we have the model parameters, we will use these as features for a supervised learning mechanism that can discriminate between normal and hypoxic conditions.

We would also like to contrast our approach to latent variable approaches, such as hidden Markov models (HMMs) and their variations. While conceptually we can think of the fetus as going through different internal states during labor, how many states there should be and how they should be connected is not known. Early attempts that we have made at learning HMM-style models have failed, because of our lack of initial model knowledge, as well as the fact that the data we have is not sufficient to fit hidden model parameters. Based on roughly 15 years of work on this problem, we strongly believe that models that are free of latent variables and that are learned from data will work successfully.

Data Description

We used a database consisting of 264 intrapartum CTG recordings for pregnancies having a birth gestational age greater than 36 weeks and having no known genetic malformations. The majority of the recordings were from normal fetuses (221 cases); the rest were severely pathological. This proportion of pathological cases was much higher than their natural incidence (Freeman, Garite, and Nageotte 2003). The normal cases were collected from a large university hospital in an urban area, which always monitors CTG during labor. The very low natural incidence of pathology necessitated collecting cases from a number of hospitals and medico-legal files. All CTG records comprised at least three hours of recording. We note that this is a larger database than other previous studies, for example, Georgoulas, Stylios, and Groumpos (2006a, 2006b) and Ozyilmaz and Yildirim (2004).

Data collection was performed by clinicians using standard clinical fetal monitors to acquire the CTG. The monitors reported at uniform sampling rates of 4 Hz for FHR (measured in beats per minute [bpm]) and 1 Hz for UP (measured in mmHg), which we up-sampled to 4 Hz by zero-insertion and low-pass filtering. In the majority of cases, the UP or FHR sensors were attached to the maternal abdomen; the FHR was acquired from an ultrasound probe and the UP was acquired by tocography (the process of recording uterine contractions during labor, using external pressure sensors attached to the maternal abdomen). In a few exceptional cases, they were acquired internally through an intrauterine (IU) probe and/or a fetal scalp electrode. Because the relationship between UP and FHR is preserved (in a qualitative sense) across different methods for measuring UP, the same modeling strategy applies to all these different types of data.

Each example was labeled by the outcome at birth, as measured both by blood gas levels and signs of neurological impairment. Preprocessing was needed to deal with loss of sensor contact, which causes a sharp drop in the signal followed by a sharp increase back to normal. We used a Schmitt trigger, which defines separate detection thresholds for down-going and up-going transitions (see Warrick et al. [2009] for details). Once dropouts are eliminated, the signal becomes a set of segments. If the dropout lasted less than 15 seconds, we used linear interpolation to reconnect these segments. Otherwise, they were left separate. Note that the data that we had to work with is particularly messy—for example some of the traces were obtained by digitizing paper printouts, rather than by saving the sensor signal directly.

System Architecture

Conceptually, the fetal heart rate can be viewed as the result of three main factors: (1) baseline heart rate (producing average cardiac output), which we model with a term f_{BL}; (2) response to maternal uterine contractions, modeled using a term f_{SI}; and (3) variability due to sympathetic-parasympathetic modulation, modeled as f_{HRV}. Consequently, we model the fetal heart rate f as the sum of three components: $f = f_{BL} + f_{SI} + f_{HRV}$. This decomposition is unique to our approach, compared to standard methods that extract FHR features. We now explain each term in more detail.

The baseline signal f_{BL} is obtained by low-pass filtering the FHR and computing a linear trend over the data window. The variability signal f_{HRV} is obtained by high-pass filtering the FHR and estimating an autoregressive model from the result.

The response to contraction f_{SI} refers to the effect that the increase in UP has on the FHR. This component of the FHR is modeled using a nonparametric linear model, based on a low-pass filtered version of the UP and the FHR with f_{BL} and f_{HRV} removed. More precisely, let these transforms of uterine pressure and fetal heart rate at time n be denoted by u_n and f_n respectively. We modeled f_n as a convolution:

$$f_n \approx \sum_{i=0}^{M-1} (h_i \Delta t) u_{n-d-i}$$

where Δt is the sampling period and h_i is a set of coefficients or parameters. From the point of view of machine learning, this is a linear model, in which the output of the system is computed by a linear combination of the inputs to the system over a history window. In signal processing, it is called an impulse response function (IRF). Two important parameters are the number of input values used in the computation, or the memory size M, and the delay d; together, d and M define the window of input signal values that will be used to estimate the output. Intuitively, for a causal system, d should be positive (that is, the output will be determined by the values of the past input). However, for our problem there is an additional measurement delay introduced by the sensor measuring the uterine pressure. Hence, since the input u is recorded by this sensor with a possible measurement delay (which depends on the quality of the sensor contact to the skin), d may be positive or negative.

For fixed values of M and d, the parameters h_i can be determined simply by least-squares estimation. However, determining the best M and d is problematic. If the system generating the data were stationary, we would expect that using more samples to estimate the output would yield lower error on the training data. However, the problem we are facing is nonstationary: the state of the fetus typically degrades over time, due to the effort of labor. This suggests that a shorter length of data should be considered, to reduce nonstationarity. To resolve this trade-off, we extracted 20-minute epochs with 10-minute overlap between successive epochs; this epoch length is much longer than the typical FHR deceleration response to a contraction (that is, 1–2 minutes). We extracted as many such epochs as possible starting from the beginning of a clean (artifact-free) segment; to include any remaining data at the end of the segment (that is, less than 10 min), the overlap was increased for the last epoch. The overlap itself was motivated by a desired to generate as much data as possible, while maintaining the correlations between epochs at a reasonable level.

Within these data restrictions, we still want to determine a good model size. The main figure of merit that we used for a model was the variance-accounted-for (VAF). Let $\mathbf{f}$ be a vector of FHR samples from a segment and $\mathbf{U}$ be the corresponding input matrix. The error is given by:

$$\mathbf{e} = \mathbf{f} - \mathbf{U}\mathbf{h}$$

The variance-accounted-for is given by:

$$\%VAF = 100 \times \left(1 - \frac{\sigma_e^2}{\sigma_f^2}\right)$$

where s^2_e and σ^2_f are the variances of the error and the output signal, respectively. Ideally, this figure should be close to 100 percent (signaling that all the variance in the signal is accounted for by the model).

Increasing M typically yields better VAF figures, but this may be due to overfitting (which is a big problem in this task, due to the amount of noise). We use two mechanisms to avoid overfitting. First, after we obtain the least-squares fit for the coefficients, we use principal component analysis (PCA) on the set of coefficients to reduce the dimensionality. Intuitively, this eliminates parameters that capture noise.

Furthermore, we need to limit the size of the memory M. To do this, we use the minimum description length (MDL) principle and add to the squared error a penalty term, proportional to the sum of the absolute differences of the consecutive coefficients. If this sum is high, the signal oscillates a lot, which is an indication of noise. To give an intuition of the effect of these choices, in figure 3 we plot, on the left, the first six principal components obtained for a particular pathological case. Note that as more components are added, the estimated output signal contains an increasing amount of high-frequency oscillations. Intuitively, this means that in the beginning we are capturing true influences, but later on we start to capture noise. The VAF continues to improve, but this improvement is marginal. Our MDL penalty increases with the amount of oscillations; using it forces the optimization to choose a lower-order model (order 2, in this case, corresponding to the point marked with a blue star), even if the higher-order models fit the observed data marginally better. Note that applying the MDL penalty results in models with different values of M for different data segments.

Once the memory parameter M is determined, the delay parameter d of the model is selected by a simple search over a range of values that were picked in an acceptable clinical range (Warrick et al. 2009).

From the clinical point of view, another important feature is the "strength" of the response of the

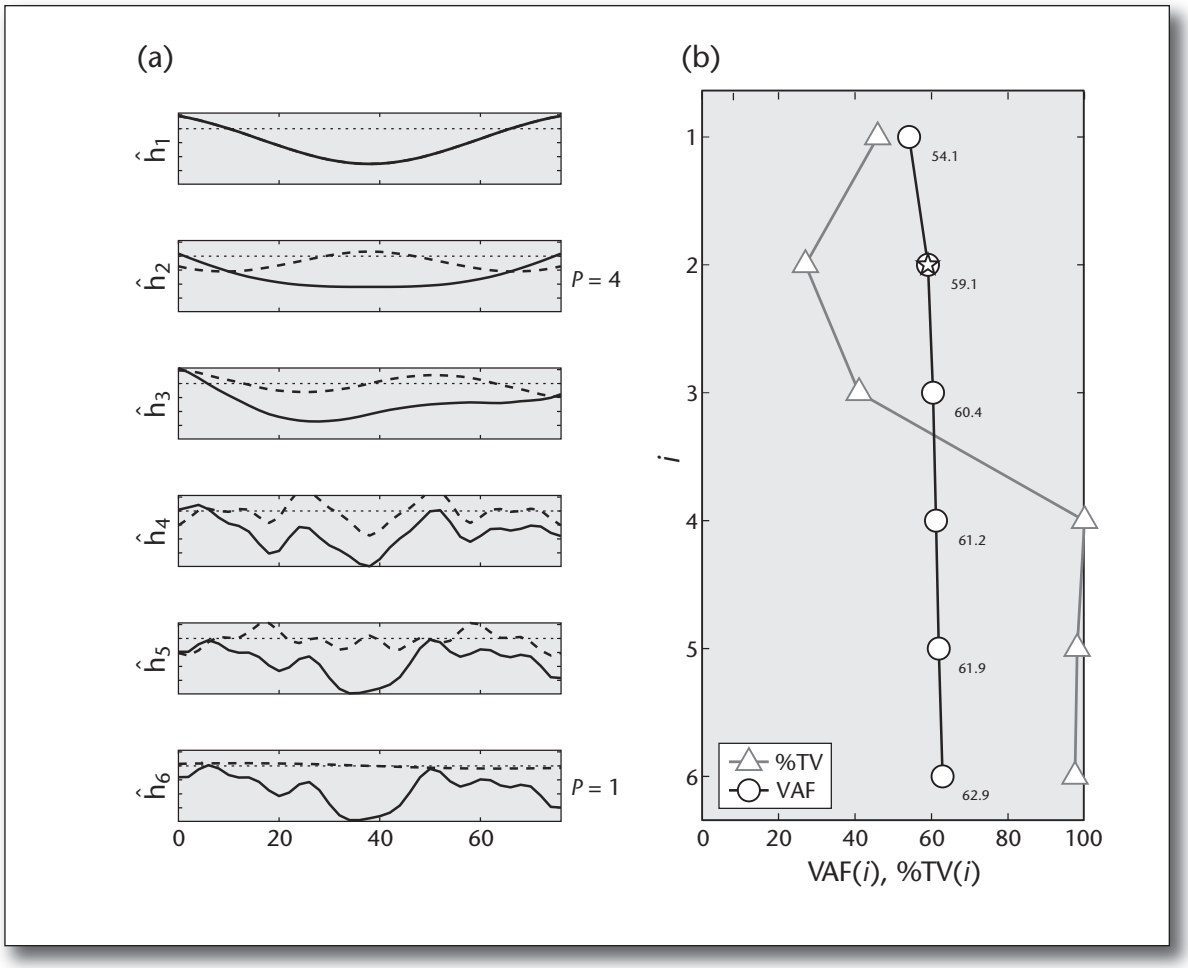

Figure 3. Order Selection for a Pathological Example.

Reproduced from Warrick et al. (2009). (a) Principal components of the IRF (dashed) and the reconstructed signal (solid) for memory length from 1 to 6. (b) VAF and penalty measure (right).

fetus to contractions. To capture this type of information, we estimate a third parameter for each model, the gain, which is the sum of the coefficients:

$$G = \sum_i h_i$$

Intuitively, the larger the gain, the stronger the response to contractions will be. If the gain is close to 0, there is almost no response to contractions. Note that the gain is well-defined regardless of the exact values of *M* and *d*. In particular, there are no statistically significant differences between the gains of models with different values of *M,* which means that the gain is a useful measure when comparing models over different data segments.

An example of data, the model, and the IRF (that is, coefficients obtained) are shown in figure 4. The FHR signal is reconstructed very well, but the high-frequency variability is not captured by this model. This is to be expected, because the contraction frequency is typically low; hence, the response to contractions must (by definition) generate a low-frequency signal.

The high-frequency content of the FHR, f_{HRV}, is clinically viewed as the result of the modulating influence of the central nervous system; in order to capture it, we high-pass filter the signal and use an autoregressive model to predict the high-passed signal. The model is also linear, but computes f_n as function of $f_n - 1 \ldots f_{n-d-M}$. We use the same MDL principle to determine the length of the model. The details are very similar to those described so far and are described in Warrick et al. (2009). Note that this high-frequency component is biologically due to the fetal nervous system, so this model captures information that is complementary to the influence of the UP.

With this setup, we are now ready to use the data set in a supervised learning setting, in order to learn how to discriminate between normal and

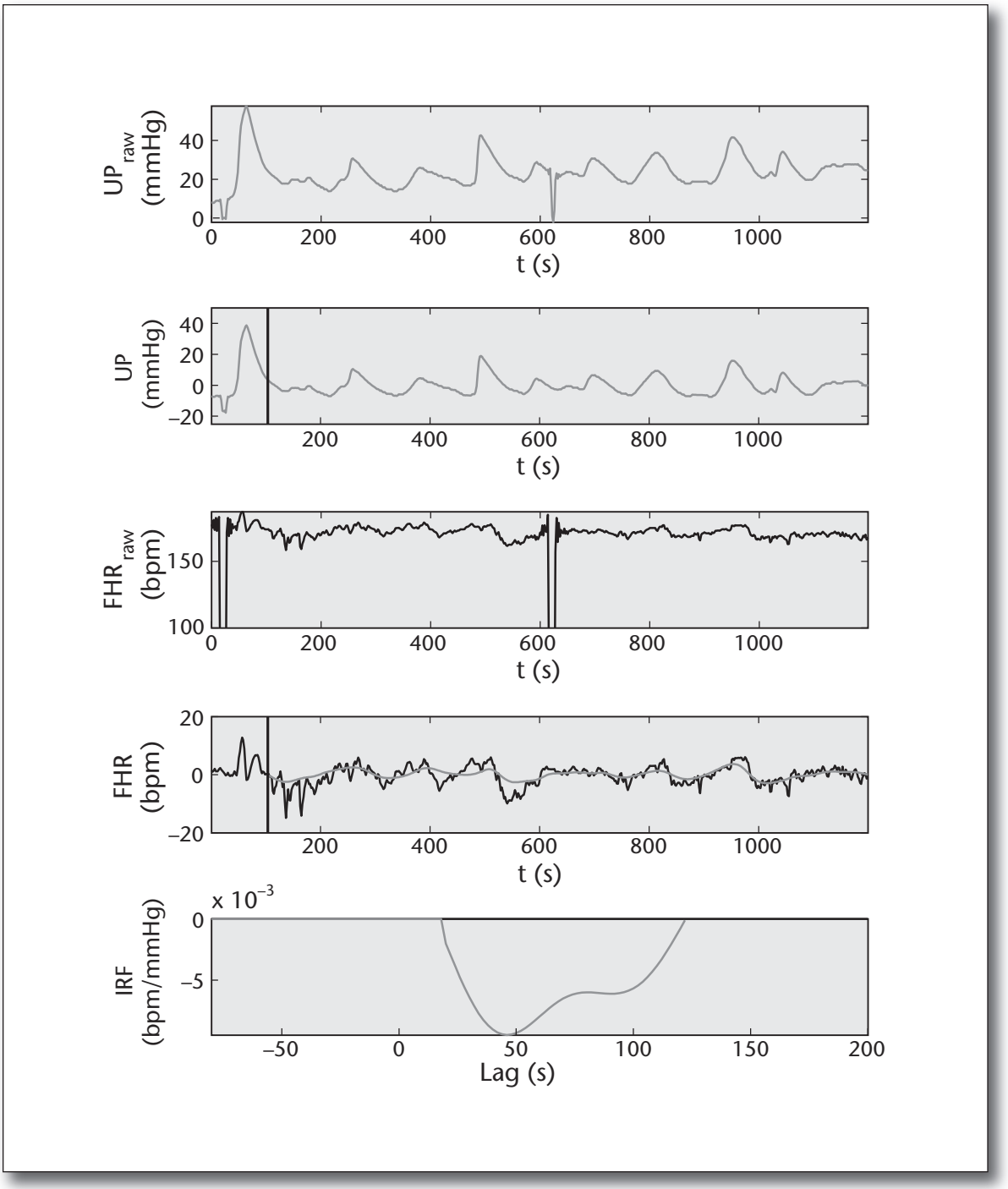

Figure 4. UP-FHR Model.

Reproduced from Warrick et al. (2009). From top to bottom: raw input UP; preprocessed UP; raw output FHR; preprocessed (black) and predicted (light gray) output FHR; final impulse response function. The IRF delay d was 20s, the gain G was –0.32 bpm/mmHg, and the VAF of the model was 44.0.

pathological babies based on their response to contractions. We labeled each segment with the fetal outcome at birth. This is an approximation, because the fetus may have started off well and degraded with time. However, this is the only reliable information available. In order to determine what model parameters to use as input to the classifier, we first did a statistical analysis to determine

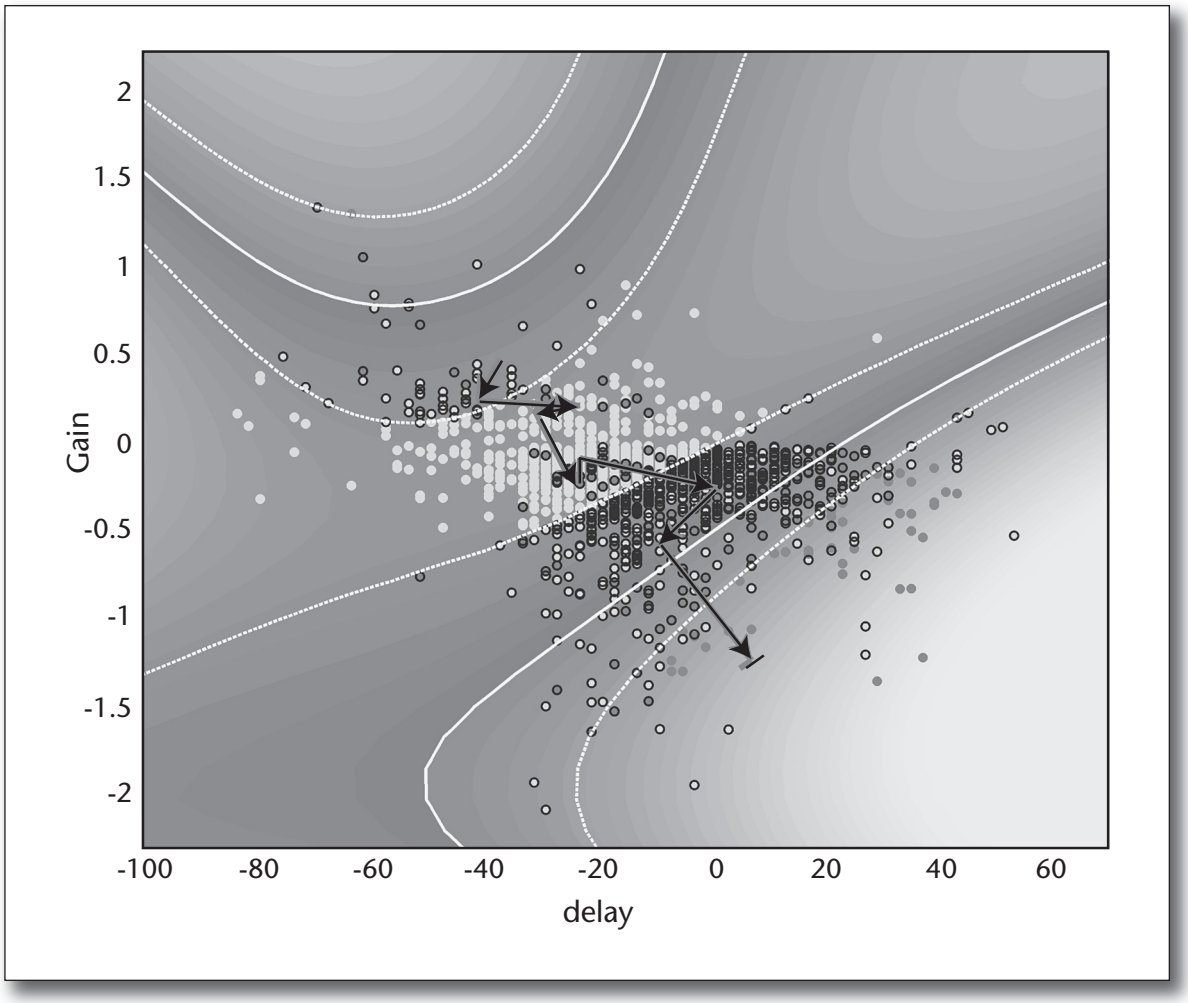

Figure 5. SVM Decision Boundary for One Fold of the System Identification Classifier.

Reproduced from Warrick et al. (2010). Normal examples are light gray and pathological ones are dark. The support vectors are outlined in black. The thick solid line represents the trajectory of a pathological case not present in the training data, which transitions from normal to pathological.

which parameters show statistically significant differences between normal and hypoxic fetuses. For these tests, parameters were considered in isolation. We found that the following parameters showed significant differences: the offset of the baseline heart rate; the gain G and the delay d, from the UP-FHR models; two measures related to the power spectrum for the heart-rate variability. We used these features as attributes for classification with support vector machines (SVMs). We used a standard SVM with a Gaussian kernel, because of its guarantees against overfitting. While the space constraints do not permit further discussion of the machine-learning methods, we refer the reader to Bishop (2006) for a detailed explanation of all methods used (SVM, PCA, and so on).

Empirical Results

We performed 10-fold cross validation, ensuring that all the data for a particular labor would be in either the training set or the test set. Note that each labor generates between 3 and 18 periods of data, so there are in effect multiple instances corresponding to each data case. It is therefore imperative to make sure that instances from the same case are not used both in the training and in the test set, so as not to bias results. All performance measures are reported on the test set.

Figure 5 shows all the instances (including the support vectors outlined in black) from one fold of training data, the learned decision function $\mathcal{H}$, and the decision boundary ($\mathcal{H} = 0$) based on the system identification feature set. The two solid

thin lines represent the class boundaries, and the dashed lines next to them show the margin around the decision boundary. We note that there are two regions in which instances are classified as pathological; the most heavily populated (lower right) is characterized by long delay and large negative gain; a smaller population in the upper left region are instances characterized by short delay and large positive gain. This roughly corresponds with the clinical observation that long, deep decelerations, as well as accelerations in response to contractions are both indicators of pathology. Our modeling and classification approach quantifies this intuition based on the available data. Between these pathological regions there is a region in which instances are classified as normal. At the boundaries of these regions are the support vectors, where classification is less certain. The large proportion of support vectors (684 of 1499 training samples) indicates that the classification problem is difficult. The trajectory of one pathological case, not included in the training set, is shown by the arrows. It began in one of the support vector regions, in which intuitively classification is somewhat uncertain, then moved into the normal region, passed through the other support vector region, and finally ended in the pathological region. This suggests that this case deteriorated from a normal to a pathological state over time. We observed other pathological cases with similar behaviour. We also observed normal cases that started normal and ended close to or within the pathological region near delivery.

The nonstationarity of the fetal state poses several challenges to the detection of pathology. Additionally, the model parameters are also nonstationary, reflecting the increasing intensity of labor (which puts stress on all babies, even healthy ones). This creates problems for per epoch classification, meaning that several instances (epochs) may be mislabeled. However, this is the best that can be done given the data we have, since the true state is not observed during labor and delivery.

We addressed this problem by introducing a detector of pathology defined by a threshold of accumulated pathological classifications. The detector avoids the confusion of decision oscillations by allowing for at most one transition, from a normal to a pathological decision. Moreover, the detector can intuitively remove some of the noise in the classifications of individual epochs. The detectors used the history of per epoch classifications for each fetus to detect pathology. Figure 6 shows the performance of the six detectors in terms of pathological detection (sensitivity) and normal detection (specificity) over time. Higher detection is better for both measures. Error bars are omitted because there was little plotting overlap and a clear ranking can be observed. Six detectors were examined using different per epoch classifiers and thresholds. Only thresholds 1 and 2 are shown because detectors with higher thresholds performed worse. It is apparent that pathological detection is more conservative (that is, delayed, as indicated by a shift to the right) for the higher threshold. We focus on C1 and C2, which use classifiers based on both features provided by the input-output model, and the baseline/heart rate variability measures. These combined detectors identified pathological cases earlier and consistently better than equivalent individual detectors. Selecting the best performing detector must consider both performance measures. We consider C2 to be the best detector because it had close to the best detection of pathological cases and close to the best false positive rates, especially in the first half of the 3-hour record, when a clinical response is most important. C2 detected half of the pathological cases with a false positive rate of 7.5 percent at epoch –10 (that is, roughly 1 hour and 40 minutes before the original time of delivery). In comparison, while C1 had that best detection of pathological cases, it had the worst false positive rates.

Discussion

The approach we proposed in this article detected correctly half of the pathological cases, with acceptable false positive rates (7.5 percent), early enough to permit clinical intervention. This detector was superior to alternatives using either feature set by itself. By definition, the pathological cases in our database had been missed by clinicians; therefore, this level of performance is quite significant. It is interesting that this corresponds well to the clinical fact that approximately 50 percent of birth-related brain injuries are deemed preventable. Timing of detection is very important given that the fetal state evolves; detecting fetal distress near the time of delivery has less potential to improve clinical outcomes, while an advance warning of 1 hour and 40 minutes is very significant clinically and can make the difference between severe injury and a positive outcome for the baby. This is a relatively long time for treatment to occur and improve outcome; typically, the interval between a decision to intervene and Cesarean birth is less than 30 minutes. Furthermore, the cost of believing these decisions (that is, a rate of unnecessary Cesarean sections of 7.5 percent) is acceptable clinically.

In order to assess the advantage of our approach compared to existing clinical practice, we recently conducted an extensive empirical comparison with a rule-based detection system designed by Parer and his colleagues (Parer and Ikeda 2007). This system was very carefully crafted and calibrated; it is considered to capture all state-of-art med-

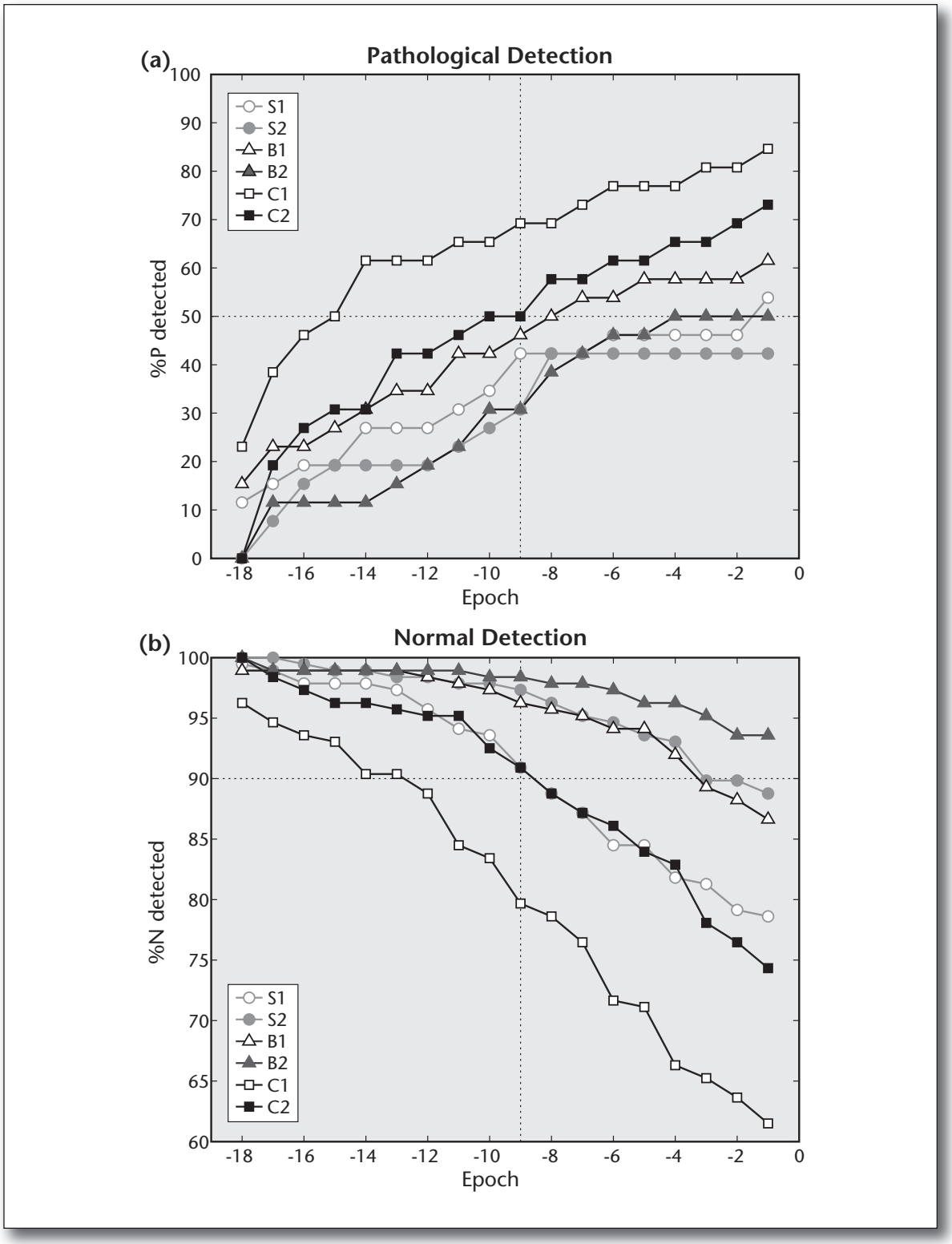

Figure 6. Pathological and Normal Detection over Time.

(a) Pathological and (b) normal detection over time for selected system identification (S1 and S2, circles), baseline-HRV (B1 and B2, triangles) and combined (C1 and C2, squares) detectors. The cumulative count is indicated by open (threshold = 1) and filled (threshold = 2) markers. The vertical dotted lines indicate the time of 90 minutes before delivery. The horizontal dotted lines indicate the 50 percent pathological and 90 percent normal detection levels. Reproduced from Warrick et al. (2010).

ical knowledge in CTG interpretation. Its detector provides a fixed sensitivity that is very close to our system, but a significantly higher false positive rate of 15 percent at 1 hour and 40 minutes before birth (Warrick et al., manuscript in preparation). This means that if the rule-based system is calibrated to provide a similar amount of advance warning for the same fraction of pathological cases as our system, it will also recommend twice the number of Cesarean sections for healthy babies. In a clinical setting with limited resources, more surgeries can divert precious resources from the patients who really need them, as well as increasing recovery times. These results obtained on our database suggest that our approach is significantly more successful. Note that the expert system against which we compare is considered at the level of human experts. Both systems therefore embody expertise not typically present in hospitals and are able to detect pathologies earlier than the physicians who happened to deliver those babies.

This is why both automated systems are able to detect the pathological cases much earlier than the birth time (and consequently, earlier than the doctors who happened to deliver those babies).

We have recently started to study a set of "intermediate" examples contained in the database, in which the oxygen level at birth was in a problematic range, but no severe pathology was detected. These cases appear to be "close calls," in which birth occurred just in time to avoid a bad outcome. A preliminary statistical analysis of the model parameters of these cases shows that they are closer to pathological cases than to normal ones. This is an encouraging result in assessing whether our system can be used to flag babies that could become problematic within a 2- to 3-hour time range. By studying these examples, we hope to understand better the timing of pathology and adjust our detection mechanism accordingly.

Making the Leap to the Clinical Setting

The approach presented in this article is currently being incorporated in PeriGen's obstetrics decision support system.[1] A unique feature of our approach, which has made it very attractive to the medical community, is the ability to use data recorded from standard clinical monitors, which would be gathered anyway during labor at any hospital. As a result, deployment does not require any new hardware; only the software system needs to be modified. Since we use signals that are already recorded, and our system is stand-alone, incorporating it into existing software has been very smooth.

Another important feature is the ability to classify data as it arrives in real time. This approach can process data directly as it arrives from the sensors. Moreover, since we do not use just a static classifier, but a detector, we can flag problems with fetal oxygenation as they arise, in a timely manner. Classifiers typically work on an entire time series, while detectors classify the data at each time step.

A crucial aspect that requires a lot of thought is the design of a successful interface between the system and the medical staff. While in this article we reported information that shows a large range of the sensitivity-specificity trade-off spectrum, in a deployed application one has to choose a particular value of sensitivity and stick with it. This choice is crucial in practice: if the system raises too many alarms, it will be turned off or ignored by medical personnel. On the other hand, the system needs to be able to detect problematic cases in a timely fashion. PeriGen has been working with a team of doctors and other medical personnel to determine how to best choose such trade-offs not only for this detector, but for their entire decision support system.

Finally, bringing these ideas to bedside requires more than just a software implementation. The technology needs to go through a lengthy approval process, which typically takes 3–4 years. The process for our system has been started, but it has yet to be completed.

Acknowledgements

This research was funded in part by the National Science and Engineering Council (NSERC) and by LMS Medical Systems Ltd. (acquired by PeriGen in 2009). We gratefully acknowledge the help of the editors of this special issue in improving the manuscript.

Note

1. See www.perigen.com/our-solutions/pericalm-EFM.

References

Bishop, C. 2006. *Pattern Classification and Machine Learning*. Berlin: Springer.

Draper, E.; Kurinczuk, J.; Lamming, C.; Clarke, M.; James, D.; and Field, D. 2002. A Confidential Enquiry into Cases of Neonatal Ecephalopathy. *Archives of Diseases in Childhood: Fetal & Neonatal* 87(3): F176–F180.

Freeman, R.; Garite, T.; and Nageotte, M. 2003. *Fetal Heart Monitoring*. Philadelphia, PA: Lippincott Williams and Wilkins.

Georgoulas, G.; Stylios, C. D.; and Groumpos, P. P. 2006a. Predicting the Risk of Metabolic Acidosis for Newborns based on Fetal Heart Rate Signal Classification Using Support Vector Machines. *IEEE Transactions on Biomedical Engineering* 55(5): 875–884.

Georgoulas, G.; Stylios, C.; and Groumpos, P. P. 2006b. Feature Extraction and Classification of Fetal Heart Rate Signals Using Wavelet Analysis and Support Vector Machines. *International Journal of Artificial Intelligent Tools* 15(3): 411–432.

Ozyilmaz, L., and Yildirim, T. 2004. ROC Analysis for

AAAI to Colocate with Cognitive Science in 2014!

AAAI is pleased to announce that it will colocate with the 2014 Cognitive Science Society Conference in picturesque Québec City, Québec, July 27-31, 2014. The conference will be held at the beautiful Centre des congrès de Québec, and attendees can stay at the adjacent Hilton Québec. More details will be available at the AAAI website soon!

Fetal Hypoxia Problem by Artificial Neural Networks. *Artificial Intelligence and Soft Computing—ICAISC 2004,* volume 3070, Lecture Notes in Artificial Intelligence, 1026–1030. Berlin: Springer.

Parer, J., and Ikeda, T. 2007. A Framework for Standardized Management of Intrapartum Fetal Heart Rate Patterns. *American Journal of Obstetrics and Gynecology* 197(1): 26–32.

Parer, J.; King, T.; Flanders, S.; Fox, M.; and Kilpatrick, S. 2006. Fetal Acidemia and Electronic Fetal Heart Rate Patterns: Is There Evidence of an Association? *Journal of Maternal-Fetal and Neonatal Medicine* 19(5): 289–294.

Saphier, C.; Thomas, E.; Brennan, D.; and Acker, D. 1998. Applying No-Fault Compensation to Obstetric Malpractice Claims. *Primary Care Update for OB/GYNS* 5(4): 208–209.

Signorini, M.; Magenes, D.; Cerutti, S.; and Arduini, D. 2003. Linear and Nonlinear Parameters for the Analysis of Fetal Heart Rate Signal from Cardiotocographic Recordings. *IEEE Transactions on Biomedical Engineering* 50(3): 365–374.

Warrick, P. 2010. Automated Decision Support for Intrapartum Fetal Surveillance. Ph.D. Dissertation, Department of Biomedical Engineering, McGill University. Montreal, QC, Canada.

Warrick, P.; Hamilton, E.; Precup, D.; and Kearney, R. 2009. Identification of the Dynamic Relationship between Intrapartum Uterine Pressure and Fetal Heart Rate for Normal and Hypoxic Fetuses. *IEEE Transactions on Biomedical Engineering* 56(6): 1587–1597.

Warrick, P.; Hamilton, E.; Precup, D.; and Kearney, R. 2010. Classification of Normal and Hypoxic Fetuses from Systems Modelling of Intra-Partum Cardiotocography. *IEEE Transactions on Biomedical Engineering* 57(4): 771–779.

Warrick, P. A.; Hamilton, E. F.; and Macieszczak, M. 2005. Neural Network Based Detection of Fetal Heart Rate Patterns. In *Proceedings of the 2005 IEEE International Join Conference on Neural Networks*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Philip Warrick received the B.A.Sc. degree in electrical engineering from the University of Waterloo in 1987, the M.Eng. degree in electrical engineering from McGill University in 1997, and the Ph.D. degree from McGill University in biomedical engineering in 2010. He has worked extensively in industry in biomedical applications, most recently (2000–2004) as a senior medical research engineer at LMS Medical Systems and at Perigen Inc. from 2009 to the present. His research interests lie in the fields of systems modeling, statistical signal processing, machine learning, and decision-support systems. His research aims to develop better methods of acquiring and interpreting biomedical signals to facilitate improved clinical decision making.

Emily F. Hamilton received the B.Sc. at Bishop's University, the MDCM at McGill University, and the FRCSC from the Royal College of Physicians and Surgeons of Canada. She is the senior vice president of clinical research at Perigen Inc. in Montreal and an adjunct professor gynecology at McGill University. An experienced obstetrician, she has also held various academic appointments at McGill University, including director of the Residency Education Program in Obstetrics and Gynecology, director of Perinatology, as well as serving on various Canadian National Task Forces defining clinical practice guidelines for fetal surveillance.

Robert Kearney received the B. Eng, M. Eng, and Ph.D from McGill University. He is a professor in the Department of Biomedical Engineering who maintains an active research program that focuses on using quantitative engineering techniques to address important biomedical problems. Specific areas of research include the development of algorithms and tools for biomedical system identification; the application of system identification to understand the role played by peripheral mechanisms in the control of posture and movement; and the development of bioinformatics tools and techniques for proteomics. Kearney is a fellow of the IEEE, the Engineering Institute of Canada, and the American Institute of Medical and Biological Engineering and a recipient of the IEEE Millennium medal.

Doina Precup received the B.Eng. degree from the Technical University of Cluj-Napoca, Cluj-Napoca, Romania, in 1994, and the M.Sc. and Ph.D. degrees from the University of Massachusetts, Amherst, in 1997 and 2000, respectively, all in computer science. Her graduate studies were funded in part by a Fulbright fellowship. She is an associate professor in the School of Computer Science, McGill University, Montreal, QC, Canada. Her current research interests include artificial intelligence, machine learning, and the application of these methods to real-world time series data.