

Neural Network Based Detection of Fetal Heart Rate Patterns

Philip Warrick^{1,3}
philip.warrick@mcgill.ca

Emily Hamilton^{2,3}
emily@lmsmedical.com

Maciej Macieszczak³
mmandb@sympatico.ca

¹Dept. of Biomedical Engineering
McGill University
Montreal, Quebec

²Dept. of Obstetrics and Gynecology
McGill University
Montreal, Quebec

³LMS Medical Systems
Medical Research and Development
Montreal, Quebec

Abstract – Automated detection of Fetal Heart Rate (FHR) patterns can potentially improve intra-partum care by providing consistent and reliable measures that assist health-care professionals in their assessment of the state of the fetus. We use the combined tools of signal processing and neural networks to detect the FHR patterns of baseline, acceleration and deceleration. Comparison to previous results reported in the literature are provided.

I. INTRODUCTION

Although childbirth is a natural process and outcomes are generally good, approximately 1-7 in 1000 babies experience sufficient oxygen deprivation during labor to cause death or brain injury (ACOG 2003, Badawi 1998, Badawi1998, Draper 2002). During labor the fetus is relatively inaccessible and the clinician must rely upon available, albeit indirect, measures of fetal condition to assess its tolerance to labor. The objective of this monitoring is to detect the fetus at substantial risk of adverse outcome so that intervention can prevent its occurrence. Today over 90% of labors are monitored electronically with sensors that measure and record fetal heart rate and maternal uterine contractions. Delay or failure to recognize abnormal patterns in these recordings can lead to fetal injury.

In fact, multiple reviews of cases with birth-related brain injury suggest that around 50% of such injuries are related to preventable medical errors, most often centering on incorrect analysis of the FHR recording (Draper 2003, Ransom 2003, Saphier 1998, Stalnacker 1997). The financial burden is massive and rising, reflecting the 4.5 million annual births in North America, the frequency of errors and the cost of an individual settlement for a baby with permanent birth related brain injury. The median jury award in single cases involving childbirth jumped 43% in one year alone, from \$700,000 in 1999 to \$1,000,000 in 2000, and continues to climb (Harming Patient Access 2003). Thus it is not surprising that childbirth healthcare services continue to generate the most frequent malpractice claims and lawsuits as well as the greatest liability exposure and cost of all medical specialties (Berry 2001).

The relationship of FHR patterns to oxygen deprivation is fairly nonspecific. The fundamental pattern is the “baseline” or resting level of the fetal heart rate. In general, a stable baseline FHR level between 120 and 160 beats per minute indicates that the baby’s heart is pumping

well and delivering adequate amounts of well-oxygenated blood. Small fluctuations around this baseline (referred to as “variability”) indicate that the central nervous system is intact and providing a healthy modulating influence. Temporary increases called “accelerations” accompany fetal movement and indicate a healthy state. Other patterns reflect adverse conditions. “Decelerations” are temporary decreases in the fetal heart rate that reflect events such as compression of the umbilical cord, malfunction of the fetal heart muscle or premature separation of the placenta. There are two main classes of decelerations based on shape and five subclasses based on duration or their temporal relationship to contractions.

Automated detection of these patterns is challenging because examples in each class can vary considerably in size and shape, reflecting diversity of the underlying process as well as the cumulative effect of several superimposed conditions. Moreover, the recordings are collected during childbirth and contain numerous gaps and artifacts related to frequent maternal and fetal movements. Finally, clinical utility requires accurate and timely detection.

II. OBJECTIVE

To develop automated techniques to detect and fetal heart rate patterns and to estimate their parameters. This must be done with a high degree of accuracy and adaptability to a real time environment.

III. METHODS

We employ a combination of ad-hoc rules, signal processing and neural-network classification to estimate FHR baseline and variability, and to detect acceleration and deceleration patterns.

A. Baseline Detection

1) Preliminary Detection of Candidate Baselines

The objective of this stage is to identify regions of the FHR tracing that may contain baselines. The extrema of the FHR are used to identify signal-enclosing boxes characterized by height, length and slope. Acceptable regions are subject to various constraints including minimum length, maximum angle and the maximum distance between any FHR value and the box midline. The

maximum distance is adaptive in that it is modulated by the ambient variability. The variability is a measure of the high frequency FHR energy ($>0.07\text{Hz}$) and is estimated by a high pass filter that is rectified and smoothed. A polynomial fitting procedure is then applied in order of priority from the longest (and most probable) regions to the shorter regions to produce linear estimates of the baseline.

2) Refinement of Candidate Baselines

Although a candidate baseline may fulfill the criteria of the polynomial approximation stage, some of these relatively straight lines are not true baselines; they can be part of a long deceleration for example. We trained a series of neural networks to recognize true and false examples of candidate baselines from 690 hours of tracings. The classification is based on the measurements of the candidate baseline in question as well as the characteristics of a set of neighboring candidate baselines. At this stage each candidate baseline is assigned a probability of correctness by the neural network and the baselines with the highest probabilities are retained.

The characteristics of a candidate baseline and its neighbouring baselines are the input features for classification: for each of the baselines in an observation window (shown in Figure 1), the features of Table 1 form the input to an ensemble of neural networks.

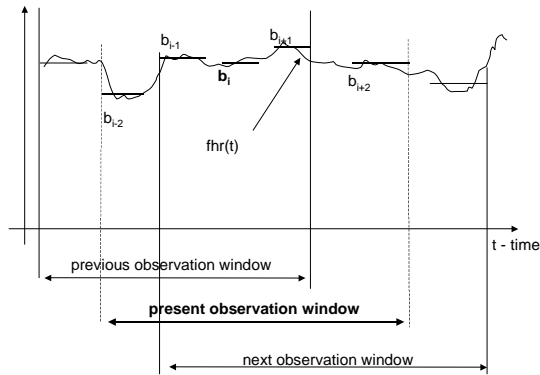


Figure 1: Observation windows for baseline refinement

Feature	Description
Length_i	length of baseline candidate b_i
YBegin_i	FHR value at the beginning of the b_i
YEnd_i	FHR value at the end of b_i
YMean_i	mean FHR level for b_i
YStd_i	Standard deviation of the FHR level for b_i
Gap_i	gap between b_i and b_{i+1}
GapMean_i	mean FHR in the gap between b_i and b_{i+1}
GapStd_i	standard deviation of the FHR in the gap between b_i and b_{i+1}

Table 1: Neural network input features for each baseline of the observation window

More generally, the measurements performed on the n baselines are used to learn the classification of the j -th baseline. All the measures are provided to the input of a neural network and the target vector is either [0] (for baseline) or [1] for non-baseline, using the labels determined by comparison with the gold standard. In the current implementation, the primary classifier learns the classification of the middle baseline ($j=3$) while the secondary classifier learns the classification of the last baseline ($j=5$). The primary classifier is not used when waiting for future baselines ($j=4,5$) because this would cause excessive delays in the system. Under these conditions the secondary classifier is used.

We use a cluster-ensemble training method that partitions the classification space into clusters identified by a specialized expert. When the system is fully trained, the complete classification space is partitioned and later recognized by a combined vote of all experts using an arbiter as a voting mechanism.

Partitioning of the classification space is a process of sequential "label-exclusion" of recognized clusters from the classification space. At the beginning the complete classification space is used to train the first expert implemented as a neural-network classifier. After training, the positively recognized cluster is "label-excluded" from the classification space by reversing its classification label (i.e. from TRUE to FALSE). After the "label-exclusion" of the recognized cluster the classification space is used to train the second expert, then the third expert and so on.

Obtaining a neural-network expert is done by a two-stage training procedure. In the first stage, a number of temporary experts are created. In the second stage all temporary experts are scanned to choose the one having the largest cluster coverage. There are two important requirements of this methodology that need to be fulfilled:

1. Every neural network should be constructed such that there is no possibility of memorization (overfitting). This can be achieved by using simple neural-network experts having sizes small enough to avoid memorization of the full classification space and large enough to learn their clusters.
2. The process of "label-exclusion" should be performed such that recognition of the cluster is done by generalization rather than memorization. In practice this simply means that a certain number of parameters that control the cluster classification should be set with values giving a recognition rate that is just above the percentage of the labeled vectors to be recognized. For example, if within the classification space there are 40% of FALSE-labels and 60% of TRUE-labels, the parameters of the given neural-network expert that

perform clustering should be set to recognized ~65-70% of TRUE-labels. And consequently those ~65-70% vectors will be then label-excluded (covered) by this expert.

After expert creation, the next step is to create a neural-network arbiter that will combine "votes" of all the trained experts into one decision. The arbiter itself is a very simple neural network that performs a weighted voting of the experts. The arbiter has a number of inputs equal to the number of experts ($N_{Experts}$) and is typically a $N_{Experts} \times 1 \times 1$ sized neural network. The value of $N_{Experts}$ used within the arbiter is chosen based on the experience gained by observing expert and arbiter trainings. We found that three experts for baseline and three experts for non-baseline recognition were sufficient. The total recognition rate of the arbiter for a given number of experts should not greatly exceed the number of initial TRUE labels within the classification space (see (2) above). Also there is a significant degradation of the cluster size during training when the number of experts exceeds the optimal value.

From our observations it appears that, given a large number of temporary experts within one cluster, the number of final experts to be chosen can have a fixed hard-coded value.

B. Acceleration-Deceleration Detection

Acceleration and deceleration events are detected in a two-step manner by first identifying candidate peaks or valleys in the signal (hereafter referred to as "bumps"), and then classifying these candidates using a neural network. While detection of accelerations is done independently from that of decelerations, the symmetry of the two problems permits a common approach to be used.

1) Bump Detection

Candidates are detected using a bank of band-pass filters to obtain events of overlapping duration ranges (from a minimum of 15s to a maximum of 5 min). The ranges are selected considering the probability distribution functions of event durations derived from expert FHR interpretation (shorter events are given tighter ranges for greater precision). After this filtering is done, the high frequency content (bumps of short duration) and the low frequency component (containing the energy of the FHR offset) are removed from the signal. The zero-crossings of this signal then delineate the event time extents (see Figure 2 for an example with one filter output). Events having insufficient area or duration are rejected using a conservative threshold that preserves sensitivity as much as possible. Events collected from each band-pass filter are then placed into a competition such that among overlapping candidates, only the highest amplitude event survives.

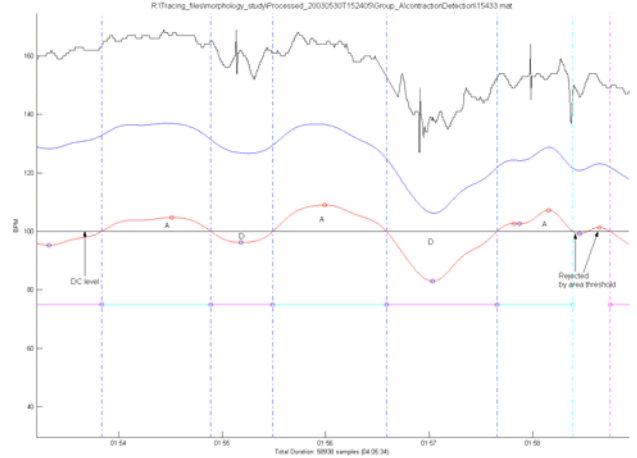


Figure 2 Candidate detection for one band-pass filter. The top and bottom signals are the original FHR and band-passed FHR, respectively. The band-passed FHR is shifted by 100 for ease of comparison. Acceleration and deceleration candidates are indicated as 'A' and 'D' respectively. The last two events are immediately rejected by the area threshold.

2) Bump Classification

To classify these candidates, an expert marked the same FHR tracings into segments of baseline, acceleration and deceleration. The decelerations were further sub-classified as either gradual or abrupt onset types. The dataset consisted of 161 FHR tracings (for a total of 762 hours) from babies having normal or neurologically impaired outcome, comprising expert markings of 5831 decelerations and 2722 accelerations. The ratio of abrupt to gradual events was roughly 3:1. The sensitivity of the candidate detector was greater than 95% for these events although it collected an additional 12438 non-deceleration and 15651 non-acceleration events.

Based on expert consultation, the features in Table 2 were deemed important to the classification task and were extracted for all candidates (see also Figure 3 and Figure 4).

These features are then applied to the training of feed-forward neural networks. Training is performed using Levenberg-Marquardt back-propagation (Hagan 1994), a second-order training algorithm, using eight-fold cross-validation. The best architecture that avoided both overfitting and underfitting was $4 \times 4 \times 2$ for accelerations and $4 \times 4 \times 3$ for decelerations. For the acceleration training the targets are either event or non-event while the deceleration targets are either gradual, abrupt or non-event. True accelerations and decelerations, being fewer in number compared to the non-events, are presented multiple times to the network (by a factor approximately equal to the event/non-event proportion) to equalize their influence during training. In cases where detected accelerations and decelerations overlapped, decelerations

took precedence, reflecting their greater clinical significance.

<i>Feature</i>	<i>Description</i>
Length	time duration of the event
Onset	time from the beginning to 90% of the peak value
Recovery	time from 90% of the peak value to the end
FhrBegin	FHR values at the beginning of the event
FhrEnd	FHR values at the end of the event
FhrStd	standard deviation of the FHR over the event
FhrInSlopeVal	steepest slope during event onset
FhrInSlopeTime	time from the beginning to FhrInSlopeVal
FhrOutSlopeVal	steepest slope during event recovery
FhrOutSlopeTime	time from the beginning to FhrOutSlopeVal
FhrPrev	FHR level in the baseline immediately preceding event
FhrNext	FHR level in the baseline immediately following event
MaxHeight	difference between the average of FhrBegin , FhrEnd and the peak FHR
MeanHeight	difference between the average of FhrBegin , FhrEnd and the mean FHR
Area	sum of differences between the mean FHR and the FHR at each event sample
VarMax	maximum variability over the event
VarPrev	minimum variability in the baseline preceding the event
VarNext	minimum variability in the baseline following the event
ContractionBegin	time elapsed since the onset of the most recent contraction
ContractionEnd	time elapsed since the end of the most recent contraction

Table 2: Acceleration and deceleration candidate features.

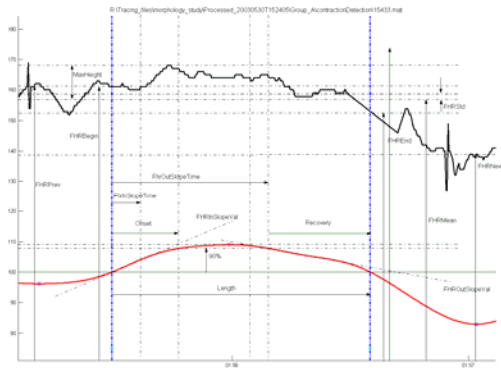


Figure 3: Measurements taken for a candidate bump. The original FHR is the top signal and the band-pass filtered version is the bottom signal.

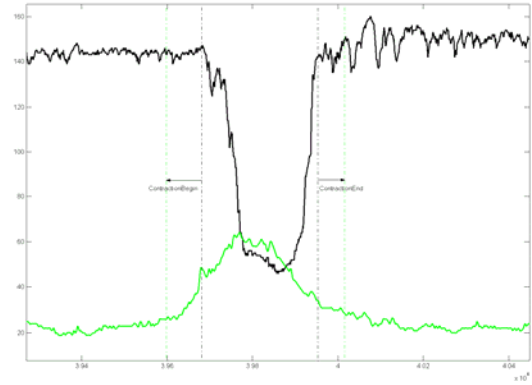


Figure 4: Measurements taken for a candidate bump for contraction proximity. The original FHR is in the top signal and the the uterine pressure is the bottom signal. The time extents of the candidate bump and it nearest contraction are shown in dashed lines.

IV. RESULTS

We created two testing standards independent from the training set. A small set of tracings were marked by five experts (Panel Standard) while a much larger set was marked by one expert (Expert Standard). The computer markings were compared to these standards.

A. Baseline

The baseline estimates were highly correlated with the baseline markings of the clinical experts in the “Standard” sets of tracings. We compared the median baseline every 15 minutes. The results are shown in Table 3 and Figure 5.

	Number of Comparisons	Correlation Coefficient	Mean Difference \pm SD
Combined standards	654	0.96	-0.9 \pm 4.2

Table 3: Detection of Baseline

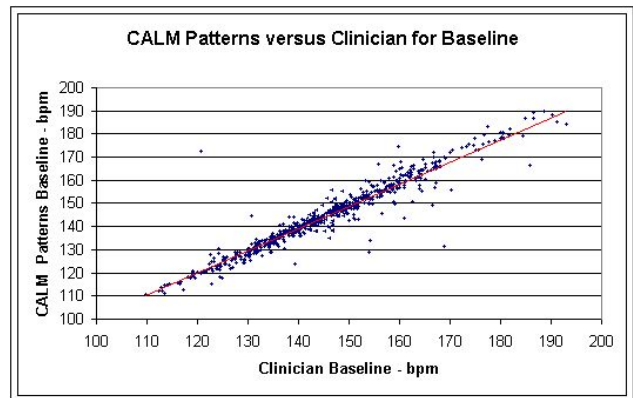


Figure 5: Comparison of CALM Patterns and clinician performance in identifying baselines

B. Accelerations and Decelerations

Table 4 summarizes the performance in the detection of accelerations and decelerations compared to the expert markings.

	P_e	TP	FP	Sensitivity	PPV
Accelerations	892	725	594	81.3%	55.0%
Decelerations	1713	1610	1215	94.0%	57.0%

Table 4: Detection of accelerations and decelerations. Sensitivity and PPV are defined as TP/P_e and $FP/(FP + TP)$ respectively, where TP is the number of true positives detected, FP is the number of false positives detected and P_e is the number of positives marked by experts.

C. Classification of decelerations according to shape

Table 5 and Table 6 summarizes the performance in the detection and classification of the abrupt and gradual deceleration types.

Abrupt Decelerations	Number	Identified as a Deceleration		Correctly Labeled as Abrupt-Onset	
		Number	%	Number	%
Combined standard	1194	1138	95.3	1013	89.0

Table 5: Detection and Labeling of Abrupt-Onset (Variable) Deceleration

Gradual Decelerations	Number	Identified as a Deceleration		Correctly Labeled as Gradual-Onset	
		Number	%	Number	%
Combined standard	519	472	90.9	396	83.9

Table 6: Detection and Labeling of Gradual-Onset (Early or Late Type) Decelerations

D. Comparison of selected Acceleration-Deceleration performance

Finally the performance of our approach to acceleration and deceleration detection is compared to recent reports in the literature in Table 7 and Table 8.

Research group	Training data		Classifier	Sensitivity (%)	Specificity (%)	PPV (%)	Aver. of PPV and Sensitivity
	Cases	hrs					
Ulbricht (1998)	'small'	-	$15 \times 3 \times 2$ recurrent NN	83.2	64.6	70.1 (*)	76.7
Ulbricht (1998)	'small'	-	Rule-based	52.0	100.0	100.0	76.0
LMS (2004)	161	762	$4 \times 4 \times 3$ MLP	94.0	80.9	57.0	75.5
Fontenla-Romero et. al. (2002)	53	58	ANFIS	70.6	75.0	76.4	73.5
Fontenla-Romero et. al. (2002)	53	58	6-11-1 MLP	66.4	62.5	66.9	66.7
Fontenla-Romero et. al. (2002)	53	58	Rule-based	68.9	27.9	52.2	60.6
Rosen et al (1996)	10	17	$300 \times 2 \times 2$ MLP	75.00	86.52	28.55	51.8
Rosen et al (1996)	10	17	C4.5	65.00	79.74	18.73	41.9
Rosen et al (1996)	10	17	Linear	59.83	86.33	23.92	41.9
Fontenla-Romero et. al. (2002)	53	58	Multi-resolution 66-6-1 MLP	83.8	78.7	-	-

Table 7: Deceleration performance comparison

Research group	Training data		Classifier	Sensitivity (%)	Specificity (%)	PPV (%)	Aver. of PPV and Sensitivity
	Cases	hrs					
Fontenla-Romero et. al. (2002)	53	58	ANFIS	72.7	58.0	65.6	69.2
LMS (2004)	161	762	$4 \times 4 \times 2$ MLP	81.3	90.4	55.0	68.2
Fontenla-Romero et. al. (2002)	53	58	Rule-based 1	63.6	66.0	67.3	65.5
Fontenla-Romero et. al. (2002)	53	58	6-11-1 MLP	56.4	60.0	61.8	59.1
Fontenla-Romero et. al. (2002)	53	58	Rule-based 2	86.5	84.5	-	-

Table 8: Acceleration performance comparison

V. DISCUSSION

Our results compare quite favourably to other reports. The ranking of the acceleration and deceleration performances by the average of sensitivity and PPV gives only an indication of the relative quality of the classifiers. The ensemble of measures must be considered in order to interpret these results. The Ulbricht rule-based approach, for example, does not seem that useful with its very low sensitivity, yet it scores highly due to its very low false positive rate. Also, given the wide disparity in training

data sizes, the confidence that can be placed in each result is likely not uniform; in this respect our results would have the highest associated confidence. Nonetheless, the composition of these results into one table does provide some overall sense of relative performance. It would be preferable to know the ROC (receiver-operator characteristic) for each of these approaches to better observe the tradeoff between sensitivity and false-positive rates and to compare methods by their area under curve (AUC). However, ROC is not generally reported and consequently the above comparison is based on single ROC samples.

While our acceleration and deceleration classifiers possess good sensitivity characteristics, both are susceptible to false positives. Based on a review of the classification results, one of the suggested reasons is a lack of event context information. In particular, in regions where events occur closely in time, some combinations are much more likely than others. The addition of more context is currently being explored using other approaches such as recursive neural networks.

To date, all studies relating FHR patterns to fetal outcome have been severely limited by visual analysis on relatively few cases. Reliable automated detection now enables the analysis of a large number of cases and the application of modern probabilistic modeling to this clinical challenge.

REFERENCES

- American College of Obstetricians and Gynecologists Task Force on Neonatal Encephalopathy and Cerebral Palsy. Neonatal Encephalopathy and Cerebral Palsy: Defining the Pathogenesis and Pathophysiology. January 2003*
- Badawi N, Kurinczuk JJ, Keogh JM, et al. Antepartum risk factors for newborn encephalopathy: the Western Australian case-control study. *BMJ* 1998;**317**:1549-1553
- Badawi N, Kurinczuk JJ, Keogh JM, et al. Intrapartum risk factors for newborn encephalopathy: the Western Australian case-control study. *BMJ* 1998;**317**:1554-1558
- Draper ES, Kurinczuk JJ, Lamming CR, et al. A confidential enquiry into cases of neonatal encephalopathy. *Arch Dis Child Fetal Neonatal ED* 2002;**87**:F176-F180
- Ransom SB, Studdert DM, Dombrowski MP, Mello JD, Brennan TA. Reduced medicolegal risk by compliance with Obstetric Clinical pathways: A case-Control Study. *Obstet Gynecol* 2003;**101**:751-5
- Stalnaker BL, Maher JE, Kleinman GE, Macksey JM, Fishman LA, Bernard JM. Characteristics of Successful claims for payment by the Florida Neurologic Injury Compensation Association Fund. *Am J Obstet Gynecol* 1997;**177**:268-71
- Saphier CJ, Thomas EJ, Studdert D, Brennan TA, Acker D. Applying no-fault compensation to obstetric malpractice claims. *Prim Care Update Ob Gyns.* 1998;**5**:208-9
- Harming Patient Access To Care: The Impact Of Excessive Litigation Statement Of The American College Of Obstetricians And Gynecologists To The Subcommittee On Health, Committee On Energy And Commerce, United States House Of Representatives, February 27 2003, www.acog.com*
- Berry G. Martin P., *Perinatal Risks. Risk Management Foundation Harvard Medical Institutions Forum: March 2001; Page 1-14*
- Hagan, M. T., Menhaj M., "Training feedforward networks with the Marquardt algorithm", *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 989-993, 1994.
- Fontenla-Romero O., Guijarro-Berdiñas B., Alonso-Betanzos A., Symbolic, neural and neuro-fuzzy approaches for pattern recognition in cardiocograms, in: H.J. Zimmermann, G. Tselentis, M. van Someren, G. Dounias (Eds.), *Advances in Computational Intelligence and Learning, Methods and Applications*, Kluwer Academic, Dordrecht, 2002.
- Guijarro-Berdiñas B., Alonso-Betanzos A., Fontenla-Romero O., Intelligent analysis and pattern recognition in cardiocographic signals using a tightly coupled hybrid system, *Artificial Intelligence*, (136) 1–27, 2002.
- Rosen B., Soriano D., Bylander T., Ortiz-Zuazaga H., Schiffrin B., Training a neural network to recognize artefacts and decelerations in cardiocograms, *1996 AAAI Spring Symposium on Artificial Intelligence in Medicine*, Stanford, CA, 1996.
- Ulbricht C., Dorffner G., A. Lee, Neural networks for recognizing patterns in cardiocograms, *Artificial Intelligence in Medicine* (12) 271–284, 1998.