Using nonlinear features for fetal heart rate classification

J. Spilka [a,*], V. Chudáček [a], M. Koucký [b], L. Lhotská [a], M. Huptych [a], P. Janků [c], G. Georgoulas [d], C. Stylios [d]

[a] Department of Cybernetics, Czech Technical University in Prague, Technicka 2, Prague 6, 166 27, Czech Republic
[b] Obstetrics and Gynaecology Clinic, 1st Medical Faculty of Charles University in Prague, Czech Republic
[c] Obstetrics and Gynaecology Clinic, University Hospital in Brno, Czech Republic
[d] Department of Informatics and Communications Technology, Technological Educational Institution of Epirus, Artas, Greece

ARTICLE INFO

ABSTRACT

• We analyzed fetal heart rate of normal and acidemic fetuses. • We used conventional and nonlinear features for the signal analysis. • Addition of nonlinear features improves accuracy of classification. • The best nonlinear features are: Lempel Ziv complexity and Sample entropy. • Combination of conventional and nonlinear features provides the best accuracy.

Abstract: Fetal heart rate (FHR) is used to evaluate fetal well-being and enables clinicians to detect ongoing hypoxia during delivery. Routine clinical evaluation of intrapartum FHR is based on macroscopic morphological features visible to the naked eye. In this paper we evaluated conventional features and compared them to the nonlinear ones in the task of intrapartum FHR classification. The experiments were performed using a database of 217 FHR records with objective annotations, i.e. pH measurement. We have proven that the addition of nonlinear features improves accuracy of classification. The best classification results were achieved using a combination of conventional and nonlinear features with sensitivity of 73.4%, specificity of 76.3%, and F-measure of 71.9%. The best selected nonlinear features were: Lempel Ziv complexity, Sample entropy, and fractal dimension estimated by Higuchi method. Since the results of automatic signal evaluation are easily reproducible, the process of FHR evaluation can become more objective and may enable clinicians to focus on additional non-cardiotocography parameters influencing the fetus during delivery.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Accurate evaluation of fetal status, based on available information, is crucial when difficulties occur during delivery. Even though a fetus is equipped with a defense mechanism to tackle the delivery induced stress, in some cases only timely intervention can prevent potential long-term consequences such as cerebral palsy, neuro-development disability, neonatal encephalopathy or even death, resulting from excessively long oxygen insufficiency [1,2].

The introduction of cardiotocography (CTG—recording of fetal heart rate (FHR) and force/pressure of contractions) in the 1960s was accompanied by great expectation since it offered a new continuous fetal surveillance method. However, meta-analysis of large multi-centric studies [3] did not prove any significant improvements in the delivery outcomes. Some studies even disproved any evidence of advantages of continuous monitoring compared to intermittent one. Moreover, CTG became the main suspect for an increased rate of cesarean sections [4]. Therefore, in 1986 CTG interpretation guidelines were introduced in order to lower the number

of asphyxiated neonates as well as the number of cesarean sections [5,6]. Nowadays CTG remains the most prevalent method for intrapartum fetal surveillance [7,8], often supported by ST-analysis method (Neoventa Medical, Sweden) which is based on analysis of fetal electrocardiogram (fECG). The introduction of additional ST-analysis into the clinical practice improved the labor outcomes slightly [9,10] but its use is not always possible or feasible since it requires invasive measurement.

A recently published paper [4] concluded that the weakness of CTG still lies in a generally poor standard of interpretation and the contribution of the human factor, demonstrated by high intra and inter observer variability. Either more education and training on CTG interpretation should be performed [1,2] or one should use a more cost-effective solution by developing a decision support system serving as a source of additional information [4].

First attempts for automatic analysis were completely based on clinical guidelines for CTG assessment [5]. Recently, other methods such as those derived from adults HRV research were used for FHR analysis [11]. The statistical description of CTG tracings was employed in [12] and in [11]. A short overview of papers which analyzed the spectrum of FHR either antepartum or intrapartum was published in [13]. An extension of frequency analysis, the wavelet transform, was employed in [14,15]. Recently, a system identi-

* Corresponding author. Tel.: +420 22435 7325.
E-mail address: spilka.jiri@fel.cvut.cz (J. Spilka).

fication approach to estimate parameters from FHR and uterine pressures was described in [16].

Use of nonlinear methods for FHR analysis has its roots in adults HRV research. The measure of fractal dimension was performed by [17–19]. Another attempt was to measure the length of FHR curve using the Higuchi method [20]; the different estimations of fractal dimension were reviewed by [21]. Probably the most successful nonlinear methods for FHR analysis are approximate entropy (ApEn) and sample entropy (SampEn). They are widely used for examination of nonlinear systems and also proved their applicability in FHR analysis [22–24]. Another method for nonlinear analysis is Lempel Ziv complexity used by [25].

For classification of the FHR many different approaches were used, e.g. Support Vector Machines [15,16,26], hybrid approach utilizing grammatical evolution [27], and also conventional methods, such as k-nn (k-nearest neighbors), ldc (linear discriminant classifier), and qdc (quadratic discriminant classifier) [26].

There are also several complete systems for fetal assessment. Among the best known are: *Omniview SisPorto*® developed at University of Porto, Portugal [28] and the NST-Expert (Non-Stress Test) and its successor *CAFE* (Computer Aided Fetal Evaluation) [29]. Both systems use a holistic approach to CTG evaluation including information about the state of the mother. Processing of the FHR in these systems is mainly based on morphological features provided by FIGO guidelines.

The purpose of our work is to distinguish between two groups: normal fetuses and fetuses with developed acidemia. The general idea of this paper is not novel; most of the features were used in papers before (cited in previous paragraph) though they were usually employed in ad hoc settings and tested on small datasets. In this paper we provide a general approach using a sufficiently large dataset and we evaluate a comprehensive set of features that originated from different domains. In addition, we consider thoroughly the applicability of nonlinear methods to FHR.

2. Data description

Data was acquired at the Dept. of Obstetrics and Gynaecology of the General University Hospital in Prague during 2007 and 2009; all women signed informed consent. The FHR signals used in this work were measured using Neoventa's STAN S21 system.

Recordings were checked for patient anamnesis and only one fold pregnancies delivered during the 38th–42nd week of pregnancy were chosen for the final database. Umbilical artery pH values were obtained to serve as objective evaluation of hypoxia. The neonatal acidemia is defined as pH below 7.05—these values were suggested in the work of Sundstrom [30]. Nevertheless there exist other works suggesting different values 7.10 [31], 7.15 [32]. Considering these facts and after consultation with obstetricians at the General University Hospital value of pH lower than 7.15 was considered not normal—further referred to as pathological.

Our database contained 217 recordings. Arterial pH values were available for all records and, based on pH threshold, 94 were considered as pathological. For comparison the expert annotations were gathered, using an in-house developed annotation system, from three experts in the field of obstetrics in the Czech Republic. Two measures were used for evaluation: intra-observer agreement as a percentage of consistently annotated records to all annotations and inter-observer agreement as a percentage of equally annotated records among the three experts to all annotations.

3. Signal preprocessing

Signals were measured either externally using Doppler ultrasound or internally by scalp electrode. Fetal heart rate recorded externally has lower signal to noise ratio than that recorded internally but there is no clinical difference between these two approaches.

Values of extracted features and further classification are highly dependent on the quality of signal preprocessing. Preprocessing steps could distort the data and add stochastic components making the use of nonlinear methods unsuitable.

In our case, the preprocessing consisted of three main steps: segment selection, artefacts removal and interpolation. An appropriate segment should be chosen as close as possible to delivery because during the last minutes major changes in fetal condition can occur. However, as it is shown in Fig. 1, FHR directly preceding the delivery is largely contaminated with artefacts and noise. Therefore, we evaluated the signal in terms of quality and automatically chose a segment with sufficient quality that is closest to delivery. Segments were 20 min long, i.e. 4800 samples using 4 Hz sampling frequency.

The FHR signal contains a lot of artefacts caused by mother and fetal movements or displacements of the transducer. In general, the amount of data being removed as artefacts or missing values is in the range between 20% and 40% of all data. The algorithm suggested by Bernades et al. [33] was used for artefact removal. Any successive five beats with a difference lower than 10 bpm among them are considered as a stable segment. Then, whenever the difference between adjacent beats is higher than 25 bpm, the sample is substituted by linear interpolation between the previous beat and the new stable segment. Thus, all abrupt changes in FHR are removed and replaced. The result of artefacts removal is presented in Fig. 2.

We used cubic Hermite spline interpolation [34], implemented in MATLAB®, to replace the missing data. We did not compute across a gap [35] when the length of the missing data was 20 s or more—the value obtained based on our experiments. For the purpose of this paper cubic Hermite spline interpolation was used for equidistant 4 Hz data re-sampling since it is convenient for signal analysis and also it is the least prone to errors when frequency spectrum is estimated [36]. The spline interpolation also introduces nonlinearity, however, the amount of nonlinearity should be approximately the same for normal and pathological FHR.

For the use of nonlinear methods signals were detrended using second order polynomial, estimated, such that uninteresting trend was removed but interesting dynamics preserved. Also we normalized signals to zero mean and unit variance.

Despite the stress caused by delivery, the FHR is also affected by fetal activity (active and quiet state) as described by [37]. Nevertheless, the fetus is mostly awake during delivery, therefore, we neglected changes in FHR caused by active–quiet states. We also neglected sex differences of FHR since we expected the difference between normal and pathological fetuses to be more profound [38].

4. Conventional features

To be able to compare our results to those of clinical praxis, we computed the morphological features introduced by the FIGO guidelines [5]. These features describe shape and changes of the FHR baseline such as baseline mean and median, number of accelerations and decelerations, long-term and short term variability, interval index, etc. Detailed description of these features is not the purpose of this paper and can be found in [11,32].

Another type of features were those used routinely in adult HRV evaluation such as NN50, RMSSD, and Poincaré plot as well as the frequency features describing the amounts of energy in different energy bins. The power spectral density was estimated using the fast Fourier transform (FFT). For more information about statistical HRV features refer to [39] and about frequency bands to [37]. Note that Poincaré plot is a nonlinear feature that is commonly
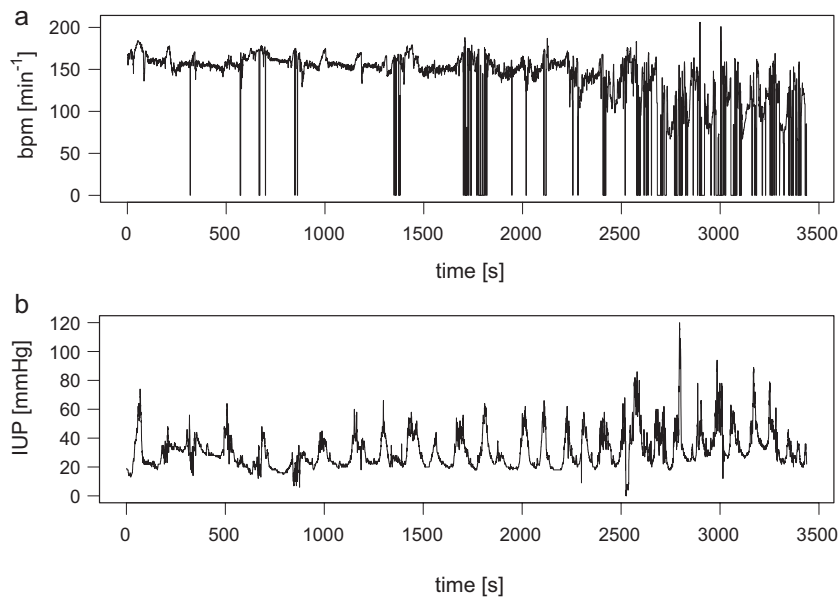
Fig. 1. Raw record of CTG: (a) fetal heart rate and (b) intrauterine pressure.

used in HRV; hence it is considered as conventional. The nonlinear features were computed over the whole 20 min segments of the preprocessed data.

5. Nonlinear features

The nonlinear approach may reveal relevant clinical information of FHR hidden to conventional time series analysis. Goldberger et al. [40] observed that a human heart beat fluctuates on different time scales and is self-similar; despite that there remains ongoing controversy over whether a normal heart rate is chaotic or not [41], tools used for examination of chaotic time series could also be useful for FHR analysis. There exist several approaches for nonlinear time series analysis; in this work fractal dimension, entropy, and complexity measures were utilized. When analyzing FHR by nonlinear methods we have to be aware of at least two major pitfalls. First, FHR contains stochastic components induced by motion

artefacts and measurement process especially when an external ultrasound electrode is used for signal acquisition. These distortions could severely damage the nature of FHR; therefore we used a surrogate data test to establish nonlinearity of FHR. Second, a certain data length is necessary to reliably estimate values of nonlinear methods. The required data length for each method is discussed in the corresponding sections below.

5.1. Fractal dimension

There are two approaches to estimate dimension of a signal either by direct measurement of the waveform or by operating in reconstructed state space. The former approach considers a signal in $\mathbb{R}^2$ as a geometric object and directly uses it without any further transformation. On the other hand the state space is reconstructed from coordinates representing the variables needed to specify the state of a dynamical system.
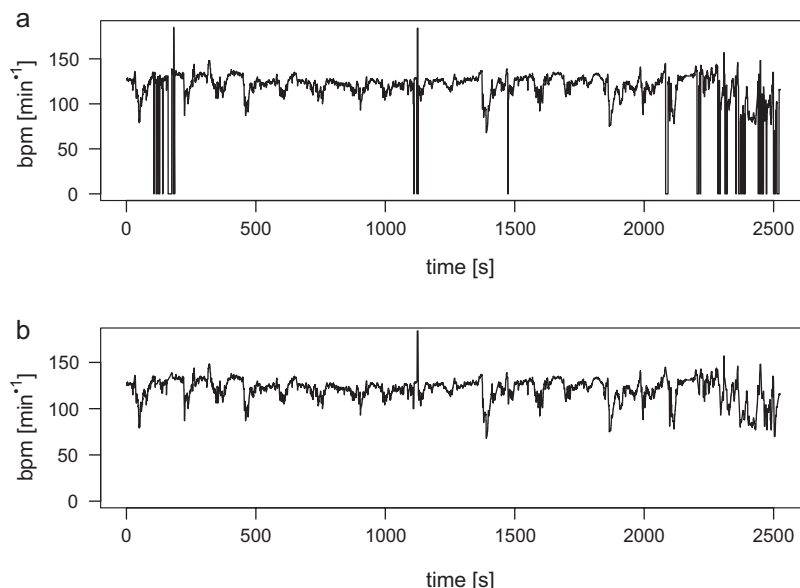
Fig. 2. Removing of artifacts: (a) raw signal with artifacts and (b) signal after artefacts removal.

a

b
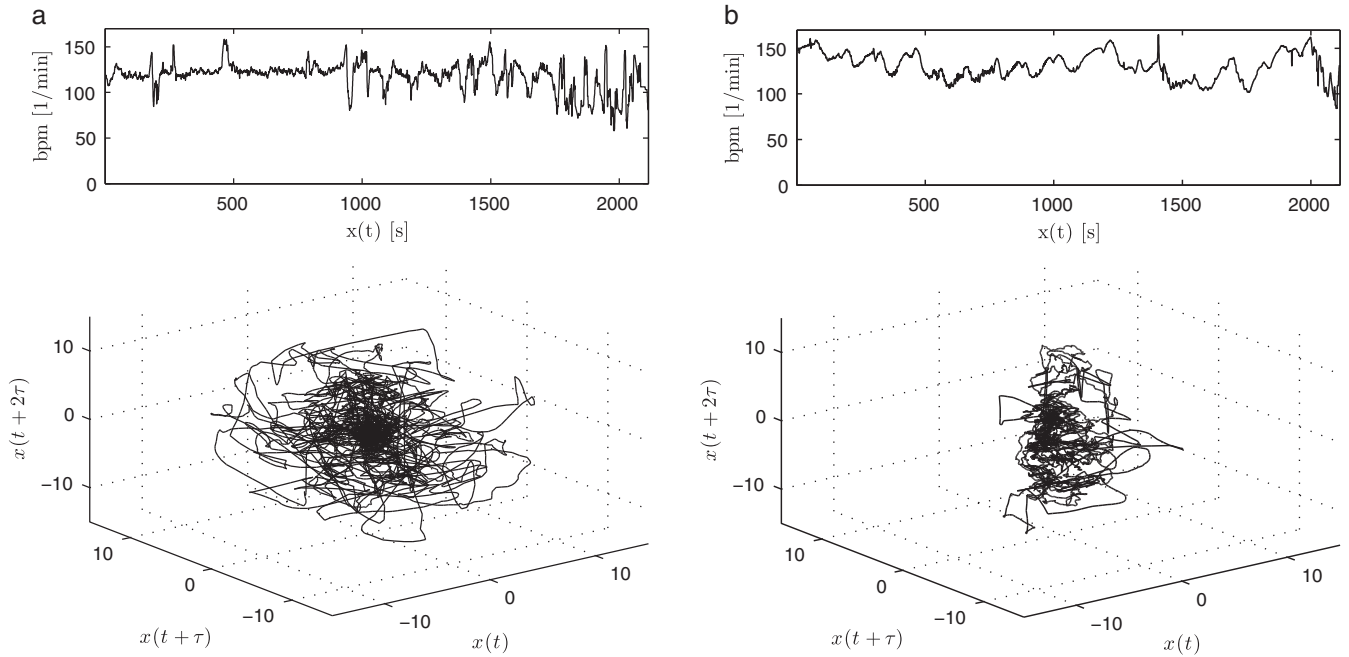
Fig. 3. Fetal heart rate and state space reconstruction for normal and pathological fetus. The upper left and right signals represent fetal heart rate. The corresponding state space is shown below. (a) FHR and state space for normal fetus. The optimal delay time was $\tau = 2.5$ s. (b) FHR and state space for pathological fetus. The optimal delay time was $\tau = 5$ s. The low complexity of FHR for fetus with developed acidemia, (b), is clearly visible in both time and state space. In the state space the delayed coordinates of FHR span less area thus showing reduced variability.

The state space can be reconstructed using Taken's embedding theorem [42]. It states that it is possible to reconstruct state space from signal $x(t)$ delayed by time τ as long as the embedding dimension m is larger than $2d + 1$, where d is a box counting dimension. $x(t) \rightarrow y(t) = [x(t), x(t+\tau), \ldots, x(t+(m-1) \cdot \tau)]$. Different choice of τ and m leads to different reconstruction. We adopted a mutual information approach [43] to search τ and considered a first marked minimum as the optimal time delay. For examination of embedding dimension we can use the Cao's method [44]. However, the correlation dimension, D_2, is invariant to m. It is therefore convenient to estimate D_2 for different m, e.g. increasing from 2 to 16, and observe when a value of D_2 saturates. This approach is computationally less efficient but minimizes errors introduced by a single m estimate. The state space reconstruction from FHR for normal and pathological fetus is shown in Fig. 3; see figure caption for details.

There are many theories about required data length with a general agreement that required data length increases exponentially with data dimension. We followed data size requirements suggested by [45]; for estimating a dimension d, a minimum data length $N_{min} = 10^{d/2}$ is required.

5.1.1. Correlation dimension

Correlation dimension D_2 is based on an estimation of the correlation sum $C(r)$ which gives the probability that two randomly chosen points are close to each other with a distance smaller than r.

$$C(m, r) = \frac{1}{N(N-1)} \sum_{i}^{N} \sum_{i>j}^{N} \Theta(r - \|\mathbf{y_i} - \mathbf{y_j}\|), \tag{1}$$

where $\mathbf{y}$ are m-dimensional delay vectors, N the number of points, and Θ is the Heaviside function: $\Theta(x)$ equals zero for $x < 0$ and one for $x \geq 0$. If we consider the limit when r approaches zero (theoretically) the correlation dimension is estimated as: $D_2 = \lim_{r \rightarrow 0} (\log C(r) / \log r)$. The D_2 is usually estimated as a slope on log–log plot where $\log C(r)$ is plotted as a function of $\log r$. Since the D_2 cannot be estimated for all radii we used local slope approach

to estimate a proper scaling region (r_l, r_u). Also we employed Takens–Theiler estimator [46,47] which uses the maximum likelihood approach:

$$D_2 = \frac{C(r_u) - C(r_l)}{\int_{r_l}^{r_u} \frac{C(r)}{r} dr} \tag{2}$$

5.1.2. Box-counting dimension

The box-counting dimension expresses the relationship between the number of boxes that contain part of an object and the size of the boxes. In the case of a signal, the minimal number of boxes N of side length ϵ, needed to cover whole signal is counted and then the side length of the boxes is decreased. The box counting dimension is estimated as follows: $D_0 = \lim_{\epsilon \rightarrow 0} (\log N / (\log(1/\epsilon)))$, though ϵ could not reach zero for real signal.

5.1.3. Higuchi's dimension

The Higuchi method [48] calculates fractal dimension from the estimated length of a signal. From an original signal $x(1)$, $x(2)$, ..., $x(N)$ of length N a new signal, X_k^m, is constructed: $X(m), X(m+k), X(m+2k), \ldots, X(m + [(N-m)/k])$ $(m = 1, 2, \ldots, k)$, where [] denotes the Gauss' notation, m defines the initial time, and k the time interval. The k represents time displacement and the number of new created subsets is equal to k. Then for each m the length $L_m(k)$ of X_k^m is computed. The length of the curve for time interval k, $\langle L(k) \rangle$, is defined as the average value over k set of $L_m(k)$:

$$\langle L(k) \rangle = \sum_{m=1}^{k} \frac{L_m(k)}{k} \tag{3}$$

The computed curve length $\langle L(k) \rangle$ for different k is related to the fractal dimension D by the exponential formula $\langle L(k) \rangle \propto k^{-D}$. The fractal dimension is estimated as a slope of fitted regression to log-log plot of $\langle L(k) \rangle$ versus k.

5.1.4. Variance dimension

The variance fractal dimension is based on properties of fractional Brownian motion (fBm). Let the signal $x(t)$ be continuous in the time t and Δt is the time increment. The variance σ^2 is related to the time increments Δt of signal $x(t)$ according to the power law [49]. $Var\{\Delta x(t_n, \Delta t)\} = \propto |\Delta t|^{2H}$, where $\Delta x(t_n, \Delta t) = x(t_n + \Delta t) - x(t_n)$ and H is the Hurst exponent computed from a log-log plot using

$$H = \lim_{\Delta t \to 0} \frac{1}{2} \frac{\log Var\{\Delta x(t_n, \Delta t)\}}{\log(\Delta t)} \tag{4}$$

Finally, the variance dimension is defined as $D_\sigma = E + 1 - H$, where E is the Euclidean dimension which is one for the time series.

5.2. Approximate entropy

Entropy describes behavior of a system in terms of randomness, and quantifies information about underlying dynamics. The approximate Entropy (ApEn) is able to distinguish low-dimensional deterministic system, chaotic system, stochastic and mixed systems [50]. It has its roots in the work of Grassberger and Procaccia [51] and Eckmann and Ruelle [52]. A time series x_n of length N is divided into a set of m-length vectors $u_m(i)$. Then the number of vectors $u_m(i)$ and $u_m(j)$, close to each other, in an Euclidean sense $d[u_m(i), u_m(j)] \le r$, is expressed by the number $n_i^m(r)$. This number is used to calculate the probability of vectors being close according to $C_i^m(r) = n_i^n/(N - m + 1)$. Let's define the function $\Phi^m(r) = 1/(N - m + 1)\sum_{i=1}^{N-m+1} \ln C_i^m(r)$. Consequently the ApEn can be defined as

$$ApEn(m, r) = \lim_{N \to \infty} [\Phi^m(r) - \Phi^{m+1}(r)]. \tag{5}$$

Generally, the tolerance r is $r = (0.15; 0.2) \cdot SD$ (SD stands for standard deviation) and the embedding dimension $m = 2$ [22]. ApEn is broadly applicable for a data series of length $N > 100$ [50]. Nevertheless, this was suggested for a wide spectrum of applications. In our case, a meaningful data length for ApEn is $N \ge 1000$.

5.3. Sample entropy

A slightly modified estimation of approximate entropy was proposed by [53] and resulted in sample entropy (SampEn). This estimation overcame the shortcomings of the ApEn mainly because the self-matches are excluded. Secondly, conditional probabilities are not estimated by a template-wise approach. SampEn requires only that one template finds a match of length $m + 1$. The calculation of SampEn is as follows:

$$SampEn(m, r) = \lim_{N \to \infty} -\ln \frac{C^{m+1}(r)}{C^m(r)}. \tag{6}$$

The values for parameters r and m were the same as for the ApEn. Also the data requirements hold.

5.4. Lempel Ziv complexity

The Lempel Ziv Complexity (LZC) [54] estimates reoccurring patterns contained in the time series irrespective of time. A periodic signal has the same reoccurring patterns and low complexity while in random signal individual patterns are rarely repeated and signal complexity is high.

A signal, $x(1), x(2), \ldots, x(n)$, is encoded to form a sequence S such that an increase in value, $x(i + 1) > x(i)$, is encoded by 1 and decrease, $x(i + 1) \le x(i)$, by 0. The algorithm counts distinct patterns in S; for each new pattern the complexity $c(n)$ is increased by 1. By convention, when the last element of S is reached, the $c(n)$ is also increased by 1. The $c(n)$ is dependent on the length of the original sequence n. We used the normalization form to avoid this dependence on

the number of data points [54]. The normalized $C(n)$ is defined as $C(n) = c(n) \cdot \log_2(n)/n$. Binary encoding was used in order to avoid a dependence of results on quantification criteria and normalization procedures. The required data length for binary encoded data is 1000 samples [55].

6. Surrogate data test

Since the FHR is a signal with finite length, finite precision, and, more importantly, is contaminated with noise, we performed a surrogate data test. The surrogate data has the same distribution and autocorrelation function as the original data but the nonlinear properties are destroyed. We used iterative Amplitude Adjusted Fourier Transform method [56] with endpoint matching to generate 39 surrogates and tested the null hypothesis that the data originates from Gaussian linear stochastic process. All methods were used as a discriminator between original data and its surrogates; the level of significance was set to $p < 0.05$.

7. Feature selection/dimensionality reduction and classification

7.1. Feature selection/dimensionality reduction

Since we acquired 33 features in total (11 FIGO-like, 14 HRV-based, and 8 nonlinear features) an automatic selection of attributes was performed. First, features were separated based on their "origin"—thus we created four sets: FIGO-like (morphological features based on FIGO-guidelines), HRV-based (features inspired by adult HRV analysis features), nonlinear (features described in Section 5), and features selected from all domains (FS-Complete). Then, based on our experience, we employed three completely unrelated feature selection/dimensionality reduction algorithms to produce a more informative input to the classifier. Therefore we employed: principal component analysis (PCA) [57]; Information gain (InfoGain) [58] – both implemented in WEKA [58] – and Group of adaptive models evolution neural network (GAME-NN) developed at the Czech Technical University in Prague [59].

The last step of the feature selection consisted of feature meta-selection, where only features selected by at least two methods were included in the final sets. The purpose of features division based on their "origin" was to prove that within each group there are features with information value and, hence, their computations were performed correctly.

7.2. Feature classification and performance evaluation

The different features sets were used to train the following classifiers: Naive Bayes, Support Vector Machine (SVM), and C4.5 decision tree; all implemented in WEKA [58]. The description of these methods could be found, e.g. in [60]. Using 10-fold cross validation the best parameters for the SVM (radial basis function kernel) were sought: the radius $\gamma = 1/2\sigma^2$ and cost parameter C.

The most common measures for the evaluation of experimental results are: sensitivity, specificity, and F-measure. The F-measure is computed using precision (= positive predictive value) $PR = TP/(TP + FP)$ and recall (= sensitivity) $RE = TP/(TP + FN)$:

$$F_\beta = \frac{(1 + \beta^2)(PR \cdot RE)}{\beta^2 \cdot PR + RE}. \tag{7}$$

Parameters TP, FP, and FN stand for true positive, false positive and false negative, respectively. The parameter β is usually set to one. This means that precision and recall have equal priority. Another useful measure is receiver operator curve (ROC) and the area under this curve (AUC). The ROC defines the relationship between sensitivity and specificity for the model parameters. Given the ROC and
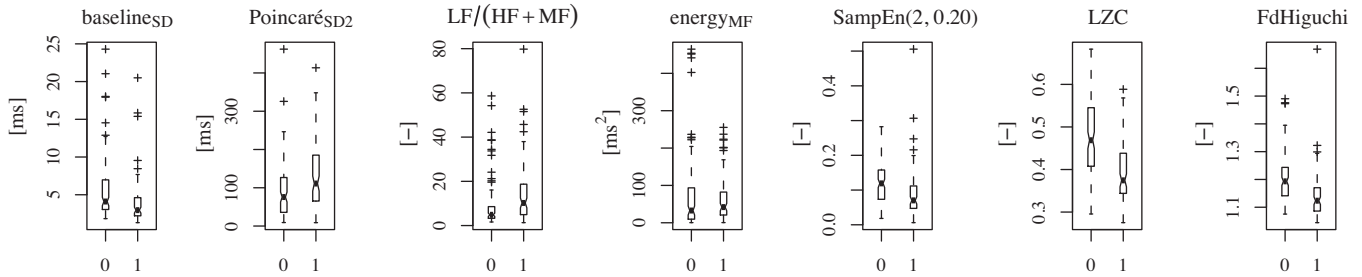
Fig. 4. Distributions of selected features: 0 – normal class and 1 – pathological class.

Table 1
Selected features organized by their origin: FIGO-like, HRV-based; nonlinear, FS-Complete.

Feature set	Selected features
FIGO-like	baselineSD, meanII
HRV-based	Poincaré_SD2, LF/(HF + MF), energyLF, energyMF, energyVLF
Nonlinear	SampEn(2,0.20), LZC, FdVariance, FdHiguchi
FS-Complete	baselineSD, Poincaré_SD2, LF/(HF + MF), energyMF, SampEn(2,0.20), LZC, FdHiguchi

AUC we can estimate which parameters are the best for modeling and if the model is suitable for the task in the first place. The experimental results were computed using 10-fold cross validation.

The difference between individual classifiers trained using different feature sets should be statistically confirmed. Although there is no unified framework the use of McNemar's test is recommended [61]. However, when dealing with a relatively small dataset (217 records in our case) there is not enough data to acceptably minimize both errors: (i) when estimating classification performance, (ii) in statistical testing. Apparently, the better way is to minimize the former error and refrain from statistical testing.

8. Results

The FHR signals were analyzed by conventional and nonlinear methods. For the correlation dimension, D_2, we found optimal time delay to be equal to $\tau = 3.9 \pm 2.5$ s. We performed the surrogate data test with all methods as discriminators and rejected the null hypotheses that data originate from Gaussian linear stochastic process on $p < 0.05$.

We performed feature selection for each group of features with the results presented in Table 1. The distributions of features from the FS-Complete set are present in Fig. 4. The discrimination between normal and pathological examples is most apparent for features: Poincaré_SD2, SampEn(2, 0.20), and LZC. The projection of the 7-dimensional feature space (FS-Complete set features) into 2-dimensional space was performed by a self organizing map [62,63] (20 × 20 units in hexagonal topology and Gaussian neighborhood trained for 200 iterations of rough training phase and 200 iterations of fine tuning phase; a sequential training algorithm was used). This map is shown in Fig. 5; the discrimination between pathological (red) and normal (green) examples is clearly visible, however there are also a number of cases that could not be easily separated. This corresponds to the classification accuracy shown in Table 3 below.

We found optimal settings for SVM (radial basis function kernel): $\gamma = 0.01$, cost parameter C = 14. Results for all features and with nonlinear features excluded are presented in Table 2. Classification results for each feature set, described in Table 1, are shown in Table 3. Results based on the expert's annotations for the same data are presented in Table 4 for comparison.

Table 2
Classification results for all features and with nonlinear features excluded.

Feature set	All in [%]	Naive Bayes	SVM	C4.5 tree
All features	Sensitivity	67.0	66.0	55.3
	Specificity	75.6	78.9	71.5
	Precision	67.7	70.5	59.8
	F-measure	67.4	68.0	57.5
	AUC	0.74	0.74	0.63
Nonlinear features excluded	Sensitivity	60.6	53.2	54.3
	Specificity	71.5	78.9	69.9
	Precision	62.0	65.8	58.0
	F-measure	61.3	58.7	56.0
	AUC	0.69	0.74	0.64

Table 3
Classification results for selected features from different groups: FIGO-like, HRV-based; nonlinear, FS-Complete.

Feature set	All in [%]	Naive Bayes	SVM	C4.5 tree
FIGO-like	Sensitivity	66.0	53.2	55.3
	Specificity	68.3	81.3	65.9
	Precision	61.4	68.5	55.3
	F-measure	63.6	64.2	55.3
	AUC	0.70	0.71	0.60
HRV-based	Sensitivity	44.7	56.4	59.6
	Specificity	82.1	83.6	69.9
	Precision	65.6	72.5	60.2
	F-measure	53.2	67.4	59.9
	AUC	0.71	0.73	0.60
Nonlinear	Sensitivity	68.9	70.1	53.3
	Specificity	73.2	78.1	85.0
	Precision	64.6	71.0	71.6
	F-measure	66.7	70.6	61.1
	AUC	0.74	0.75	0.68
FS-Complete	Sensitivity	72.3	**73.4**	60.6
	Specificity	75.6	**76.3**	71.5
	Precision	69.4	**70.3**	62.0
	F-measure	70.8	**71.9**	61.3
	AUC	0.75	**0.78**	0.68

When the nonlinear features were excluded and only conventional features were used the classification performance decreased as is shown in Table 2. The best results were achieved using the SVM with FS-Complete feature set, shown in bold in Table 3.

Table 4
Classification results of clinicians when comparing their evaluation of FHR to objective pH annotation.

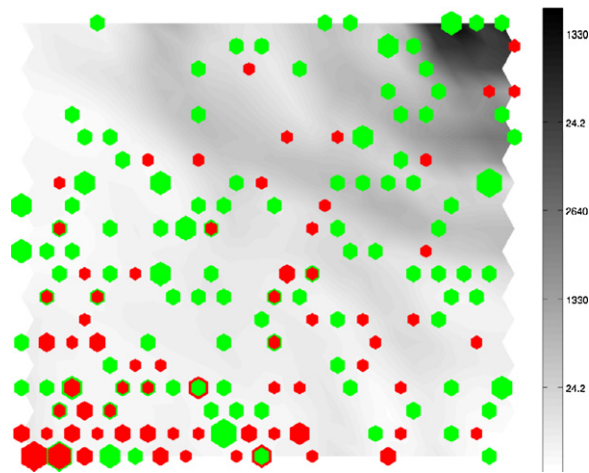| All in [%]	Expert #1	Expert #2	Expert #3
Sensitivity pH	34.4	49.0	40.5
Specificity pH	14.1	16.2	8.6
Intra-observer agreement	70.7	56.1	76.7
Inter-observer agreement between the three experts		80.5	

Fig. 5. Projection of 7-dimensional FS-Complete feature set using self organizing map into 2-dimensions. Green – normal class, red – pathological class, grey scale represents distance between examples.

9. Discussion

When analyzing the FHR one has to be aware of its properties especially when using nonlinear methods. As an example we can use correlation dimension D_2. The values of D_2 for all records $D_2 = 2.42 \pm 0.61$ were low in comparison to $D_2 \approx 6$ of adult HRV [64]. There are two possible explanations. First, the RR intervals from external (ultrasound) records were corrupted by the measurement process, such as quantization, filtering, and averaging. To be more specific, the external monitoring uses Doppler ultrasound for fetal heart beat detection which are usually determined from a periodicity of Doppler envelope using autocorrelation function. This function tends to average slight successive changes of heart beats resulting in loss of FHR variability [65]. Second, during the preprocessing stage, artefacts were linearly interpolated, i.e. a possible complex behavior was replaced by a line. Both factors may contribute to decreased value of D_2. Nevertheless, the nonlinear methods are still applicable since the surrogate date test showed significant difference between FHR and its surrogates on $p < 0.05$. However, there is also a limitation to the surrogate test due to nonlinearity introduced by the cubic spline interpolation.

Regarding the classification results there are two important questions that need to be carefully considered. How to interpret the results acquired and presented in Tables 2–4? How to relate them to the works already done? The answer is not straightforward since there are two quite distinct approaches to processing and classification of the FHR signal intrapartum.

The first one is a more technical approach that uses the objective evaluation of the data. This approach was used in many papers during the last decade and was also adopted in this paper. Our results of 73.4% sensitivity, 76.3% of specificity, and F-measure 71.9% compare favourably to those of [26,66] especially considering the small sample size they used. However, this technical approach suffers from, at least, two major drawbacks. The relation of hypoxia to the fetal cord arterial pH after delivery is widely discussed in several papers [31]. The predominant conclusion is that only an overall examination of the baby at about four years of age can bring a confident enough conclusion on the occurrence of effective asphyxia during the delivery. In addition, in many cases where timely interventions based on suspect/pathological FHR signal is made, the arterial pH values of the instrumentally delivered baby will be above the pathological threshold.

A second approach to the evaluation of FHR recordings is to acquire expert evaluation of signals and use the classification pro-

cess to try to adopt an expert behavior. Nevertheless, this approach has several drawbacks as well. First, the inter and even the intra observer variabilities are quite substantial as presented in Table 4. Second, the experts categorize the signals usually according to FIGO-based guidelines into the three classes (normal, suspect, and pathological). Large subset of signals are evaluated as suspect, but suspect class does not exist after delivery, there is usually normal or (possibly) asphyxiated baby (about whom, there will likely not be any decisive proof for at least the next several months). Moreover, in the clinical practice there is always additional information involved in the decision making process—pushing the decision process more in one direction.

10. Conclusion

This paper, for the first time, evaluates the behavior of the full set of nonlinear methods on one reasonably large database. In addition we considered the usability of these methods with respect to FHR properties. We compared a full spectrum of nonlinear features to the conventional features already used in other papers, and we showed that classification based on feature sets including nonlinear features performs the best. It also performs on-par with the best results obtained in other works such as [26,66].

The evaluation of FHR still remains subjective [67]. Clinicians say that they use conventional features like accelerations, decelerations, and variability though there is a contributing factor of pattern-like memory acquired during working experience. In contrast to the previous works, we proved that nonlinear features are useful in combination with conventional ones on a large database.

To be able to affect the way the data is processed in the clinical settings there is a need to include additional information describing the environment in which the FHR signal is acquired (maternal preexisting condition, drugs used, length of delivery, etc.) which are an integral part of the obstetrician's decision making progress. Nevertheless, it is evident that proper processing of the FHR will remain an important part of any future decision support system.

References

[1] V. Doria, A.T. Papageorghiou, A. Gustafsson, A. Ugwumadu, K. Farrer, S. Arulkumaran, Review of the first 1502 cases of ECG-ST waveform analysis during labour in a teaching hospital, BJOG 114 (2007) 1202–1207.
[2] M. Westerhuis, A. Kwee, A.A. van Ginkel, A.P. Drogtrop, W.J.A. Gyselaers, G.H.A. Visser, Limitations of ST analysis in clinical practice: three cases of intrapartum metabolic acidosis, BJOG 114 (2007) 1194–1201.
[3] Z. Alfirevic, D. Devane, G.M.L. Gyte, Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour, Cochrane Database Syst. Rev. 3 (2006) CD006066.
[4] P.J. Steer, Has electronic fetal heart rate monitoring made a difference, Semin. Fetal Neonat. Med. 13 (2008) 2–7.
[5] FIGO, Guidelines for the use of fetal monitoring, Int. J. Gynecol. Obstet. 25 (1986) 159–167.
[6] NIH, Electronic fetal heart rate monitoring: research guidelines for interpretation. National Institute of Child Health and Human Development Research Planning Workshop, Am. J. Obstet. Gynecol. 177 (1997) 1385–1390.
[7] J. Bernardes, A. Costa-Pereira, D.A. de Campos, H.P. van Geijn, L. Pereira-Leite, Evaluation of interobserver agreement of cardiotocograms, Int. J. Gynaecol. Obstet. 57 (1997) 33–37.
[8] O. Palomäki, T. Luukkaala, R. Luoto, R. Tuimala, Intrapartum cardiotocography—the dilemma of interpretational variation, J. Perinat. Med. 34 (2006) 298–302.
[9] H. Norén, I. Amer-Wåhlin, H. Hagberg, A. Herbst, I. Kjellmer, K. Maršál, P. Olofsson, K.G. Rosén, Fetal electrocardiography in labor and neonatal outcome: data

from the Swedish randomized controlled trial on intrapartum fetal monitoring, Am. J. Obstet. Gynecol. 188 (2003) 183–192.

[10] I. Amer-Wåhlin, K. Maršál, ST analysis of fetal electrocardiography in labor, Semin. Fetal Neonat. Med. 16 (2011) 29–35.

[11] H. Gonçalves, A.P. Rocha, D.A. de Campos, J. Bernardes, Linear and nonlinear fetal heart rate analysis of normal and acidemic fetuses in the minutes preceding delivery, Med. Biol. Eng. Comput. 44 (2006) 847–855.

[12] G. Magenes, M.G. Signorini, D. Arduini, Classification of cardiotocographic records by neural networks, in: Proc. IEEE-INNS-ENNS International Joint Conference on Neural Networks IJCNN 2000, 3, 2000, pp. 637–641.

[13] J. Van Laar, M. Porath, C. Peters, S. Oei, Spectral analysis of fetal heart rate variability for fetal surveillance: review of the literature, Acta Obstet. Gynecol. Scand. 87 (2008) 300–306.

[14] E. Salamalekis, P. Thomopoulos, D. Giannaris, I. Salloum, G. Vasios, A. Prentza, D. Koutsouris, Computerised intrapartum diagnosis of fetal hypoxia based on fetal heart rate monitoring and fetal pulse oximetry recordings utilising wavelet analysis and neural networks, BJOG 109 (2002) 1137–1142.

[15] G. Georgoulas, C.D. Stylios, P.P. Groumpos, Feature extraction and classification of fetal heart rate using wavelet analysis and support vector machines, Int. J. Artif. Intell. Tools 15 (2005) 411–432.

[16] P. Warrick, E. Hamilton, D. Precup, R. Kearney, Classification of normal and hypoxic fetuses from systems modeling of intrapartum cardiotocography, IEEE Trans. Biomed. Eng. 57 (2010) 771–779.

[17] D.G. Chaffin, C.C. Goldberg, K.L. Reed, The dimension of chaos in the fetal heart rate, Am. J. Obstet. Gynecol. 165 (1991) 1425–1429.

[18] N.A. Gough, Fractal analysis of foetal heart rate variability, Physiol. Meas. 14 (1993) 309–315.

[19] C.S. Felgueiras, J.P. de Sá, J. Bernardes, S. Gama, Classification of foetal heart rate sequences based on fractal features, Med. Biol. Eng. Comput. 36 (1998) 197–201.

[20] A. Kikuchi, N. Unno, T. Horikoshi, T. Shimizu, S. Kozuma, Y. Taketani, Changes in fractal features of fetal heart rate during pregnancy, Early Hum. Dev. 81 (2005) 655–661.

[21] P. Hopkins, N. Outram, N. Lofgren, E.C. Ifeachor, K.G. Rosen, A comparative study of fetal heart rate variability analysis techniques, in: Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society —Proceedings, 2006, pp. 1784–1787.

[22] S.M. Pincus, R.R. Viscarello, Approximate entropy: a regularity measure for fetal heart rate analysis, Obstet. Gynecol. 79 (1992) 249–255.

[23] M. Ferrario, M.G. Signorini, G. Magenes, S. Cerutti, Comparison of entropy-based regularity estimators: application to the fetal heart rate signal for the identification of fetal distress, IEEE Trans. Biomed. Eng. 53 (2006) 119–125.

[24] H. Gonçalves, J. Bernardes, A.P. Rocha, D.A. de Campos, Linear and nonlinear analysis of heart rate patterns associated with fetal behavioral states in the antepartum period, Early Hum. Dev. 83 (2007) 585–591.

[25] M. Ferrario, M. Signorini, G. Magenes, Complexity analysis of the fetal heart rate variability: early identification of severe intrauterine growth-restricted fetuses, Med. Biol. Eng. Comput. 47 (2009) 911–919.

[26] G. Georgoulas, C.D. Stylios, P.P. Groumpos, Predicting the risk of metabolic acidosis for newborns based on fetal heart rate signal classification using support vector machines, IEEE Trans. Biomed. Eng. 53 (2006) 875–884.

[27] G. Georgoulas, D. Gavrilis, I.G. Tsoulos, C.D. Stylios, J. Bernardes, P.P. Groumpos, Novel approach for fetal heart rate classification introducing grammatical evolution, Biomed. Signal Process. Control 2 (2007) 69–79.

[28] D.A. de Campos, P. Sousa, A. Costa, J. Bernardes, Omniview-SisPorto® 3.5—a central fetal monitoring station with online alerts based on computerized cardiotocogram + ST event analysis, J. Perinat. Med. 36 (2008) 260–264.

[29] B. Guijarro-Berdinas, A. Alonso-Betanzos, Empirical evaluation of a hybrid intelligent monitoring system using different measures of effectiveness, Artif. Intell. Med. 24 (2002) 71–96.

[30] A. Sundström, D. Rosén, Fetal Surveillance—Textbook [online], Neoventa Medical AB, Gothenburg, Sweden, 2000.

[31] H. Cao, D.E. Lake, I. Ferguson, C.A. Chisholm, M.P. Griffin, J.R. Moorman J.E., Toward quantitative fetal heart rate monitoring, IEEE Trans. Biomed. Eng. 53 (2006) 111–118.

[32] G. Georgoulas, C. Stylios, G. Nokas, P. Groumpos, Classification of fetal heart rate during labour using hidden Markov models, in: Proceedings of the 2004 IEEE International Joint Conference on Neural Networks, 3, 2004, pp. 2471–2475.

[33] J. Bernardes, C. Moura, J.P. de Sa, L.P. Leite, The Porto system for automated cardiotocographic signal analysis, J. Perinat. Med. 19 (1991) 61–65.

[34] D.K. Kahaner, C. Moler, S. Nash, Numerical Methods and Software, Prentice-Hall, 1989.

[35] J.C. Sprott, Chaos and Time-Series Analysis, Oxford University Press, 2003.

[36] M. Cesarelli, M. Romano, M. Ruffo, P. Bifulco, G. Pasquariello, A. Fratini, PSD modifications of FHRV due to interpolation and CTG storage rate, Biomedical Signal Processing and Control (2010), doi:10.1016/j.bspc.2010.10.002.

[37] M.G. Signorini, G. Magenes, S. Cerutti, D. Arduini, Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings, IEEE Trans. Biomed. Eng. 50 (2003) 365–374.

[38] J. Bernardes, H. Gonçalves, D. Ayres-De-Campos, A. Rocha, Sex differences in linear and complex fetal heart rate dynamics of normal and acidemic fetuses in the minutes preceding delivery, J. Perinat. Med. 37 (2009) 168–176.

[39] M. Malik, A.J. Camm, J.T. Bigger Jr., G. Breithardt, S. Cerutti, R.J. Cohen, P. Coumel, E.L. Fallen, H.L. Kennedy, R.E. Kleiger, F. Lombardi, A. Malliani, A.J. Moss, J.N. Rottman, G. Schmidt, P.J. Schwartz, D.H. Singer, Heart rate variability. Standards of measurement, physiological interpretation, and clinical use, European Heart Journal 17 (3) (1996) 354–381.

[40] A. Goldberger, V. Bhargava, B. West, A. Mandell, On the mechanism of cardiac electrical stability, the fractal hypothesis, Biophys. J. 48 (1985) 525–528.

[41] L. Glass, Introduction to controversial topics in nonlinear science: is the normal heart rate chaotic? Chaos 19 (2009) 028501.

[42] F. Takens, Detecting strange attractors in turbulence, Dynam. Syst. Turbulence 4 (1981) 366–381.

[43] A.M. Fraser, H.L. Swinney, Independent coordinates for strange attractors from mutual information, Phys. Rev. A 33 (2) (1986) 1134–1140.

[44] L. Cao, Practical method for determining the minimum embedding dimension of a scalar time series, Physica D 110 (1997) 43–50.

[45] D. Ruelle, Deterministic chaos: the science and the fiction, Proc. R. Soc. Lond. 427 (1990) 241–248.

[46] F. Takens, On the numerical determination of the dimension of an attractor, Dynam. Syst. Bifurcat. 1125 (1985) 99–106.

[47] J. Theiler, Lacunarity in a best estimator of fractal dimension, Phys. Lett. A 133 (1988) 195–200.

[48] T. Higuchi, Approach to an irregular time series on the basis of the fractal theory, Physica D 31 (1988) 277–283.

[49] W. Kinsner, Batch and real-time computation of a fractal dimension based on variance of a time series, Technical Report, Department of Electrical & Computer Engineering, University of Manitoba, Winnipeg, Canada, 1994.

[50] S. Pincus, Approximate entropy (ApEn) as a complexity measure, Chaos 5 (1) (1995) 110–117.

[51] P. Grassberger, I. Procaccia, Measuring the strangeness of strange attractors, Physica D: Nonlinear Phenom. 9 (1983) 189–208.

[52] J.P. Eckmann, D. Ruelle, Ergodic theory of chaos and strange attractors, Rev. Mod. Phys. 57 (1985) 617–656.

[53] J.S. Richman, J.R. Moorman, Physiological time-series analysis using approximate entropy and sample entropy, Am. J. Physiol. Heart Circ. Physiol. 278 (2000) H2039–H2049.

[54] A. Lempel, J. Ziv, On the complexity of finite sequences, IEEE Trans. Inform. Theory IT-22 (1) (1976) 75–81.

[55] M. Ferrario, M.G. Signorini, S. Cerutti, Complexity analysis of 24 hours heart rate variability time series, Conf. Proc. IEEE Eng. Med. Biol. Soc. 6 (2004) 3956–3959.

[56] T. Schreiber, A. Schmitz, Improved surrogate data for nonlinearity tests, Phys. Rev. Lett. 77 (4) (1996) 635–638.

[57] C. Bishop, Neural networks for pattern recognition, Clarendon Press, Oxford, 1995.

[58] I.H. Witten, E. Frank, Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, San Francisco, 2005.

[59] P. Kordik, Fully Automated Knowledge Extraction using Group of Adaptive Models Evolution, Ph.D. thesis, Czech Technical University in Prague, FEE, Dep. of Comp. Sci. and Computers, FEE, CTU Prague, Czech Republic, 2006.

[60] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd Edition, Wiley-Interscience, 2000.

[61] S. Salzberg, On comparing classifiers: pitfalls to avoid and a recommended approach, Data Min. Knowl. Discov. 1 (1997) 317–328.

[62] T. Kohonen, Self-Organizing Maps, volume 30, Series in Information Sciences, Springer, Heidelberg, 2001.

[63] J. Vesanto, J. Himberg, E. Alhoniemi, J. Parhankangas, SOM Toolbox for Matlab 5, Helsinki University of Technology, Report A57, Libella Oy Espoo, 2000.

[64] P. Van Leeuwen, H. Bettermann, U. An der Heiden, H. Kummell, Circadian aspects of apparent correlation dimension in human heart rate dynamics, Am. J. Physiol. - Heart Circ. Physiol. 269 (1995) 130–134.

[65] J. Jezewski, J. Wrobel, K. Horoba, Comparison of Doppler ultrasound and direct electrocardiography acquisition techniques for quantification of fetal heart rate variability, IEEE Trans. Biomed. Eng. 53 (2006) 855–864.

[66] T.K. Chung, M.P. Mohajer, Z.J. Yang, A.M. Chang, D.S. Sahota, The prediction of fetal acidosis at birth by computerised analysis of intrapartum cardiotocography, Br. J. Obstet. Gynaecol. 102 (1995) 454–460.

[67] I. Amer-Wåhlin, K. Källén, A. Herbst, H. Rydhstroem, A. Sundström, K. Maršál, Implementation of new medical techniques: experience from the Swedish randomized controlled trial on fetal ECG during labor, J. Matern. Fetal Neonat. Med. 18 (2005) 93–100.