Using nonlinear features for fetal heart rate classification

J. Spilka\textsuperscript{a,}\textsuperscript{*}, V. Chud\u{a}\v{c}ek\textsuperscript{a}, M. Koucky\textsuperscript{b}, L. Lhotsk\u{a}\textsuperscript{a}, M. Huptych\textsuperscript{a}, P. Jank\u{u}\textsuperscript{c}, G. Georgoulas\textsuperscript{d}, C. Stylios\textsuperscript{d}

\textsuperscript{a}Department of Cybernetics, Czech Technical University in Prague, Technicka 2, Prague 6, 166 27
\textsuperscript{b}Obstetrics and Gynaecology clinic, 1\textsuperscript{st} Medical Faculty of Charles University in Prague, Czech Republic
\textsuperscript{c}Obstetrics and Gynaecology clinic, University Hospital in Brno, Czech Republic
\textsuperscript{d}Department of Informatics and Communications Technology, Technological Educational Institution of Epirus, Arta, Greece

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.

Keywords: fetal heart rate, cardiotocography, nonlinear methods, feature selection, classification.

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 1960’s 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 identification 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, 18, 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, 23, 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 (Non-

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. Pre-

processing 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 de-

livery because during the last minutes major changes in fetal condition can occur. However, as it is shown in Figure 1, FHR directly preceeding 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 minutes long, i.e. 4800 samples using 4 Hz sampling frequency.

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 hy-

poxia. 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 consulta-

tion 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 anno-

tation 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 anno-

tated records to all annotations and inter-observer agreement as a percentage of equally annotated records among the three ex-

perts to all annotations.
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 seconds 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.

Figure 2: Removing of artefacts. (a) Raw signal with artefacts, (b) signal after artefacts removal.

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 [38]. 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 used in HRV; hence it is considered as conventional. The nonlinear features were computed over the whole 20 minutes 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.

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 \( \tau \) 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 \( \tau \) and \( m \) leads to different reconstruction. We adopted a mutual information approach [43] to search \( \tau \) 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 Figure 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 \( N_{\text{min}} = 10^{7}/2 \) 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 \( r \).

\[
C(m, r) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j>i}^{N} \Theta(r - ||y_i - y_j||).
\]

where \( y \) are \( m \)-dimensional delay vectors, \( N \) the number of points, and \( \Theta \) 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: \( D_2 = \lim_{r \to 0} \frac{\log C(r)}{\log r} \). 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:

\[
D_2 = \frac{C(r_u) - C(r_l)}{\int_{r_l}^{r_u} \frac{\log C(r)}{r} dr}
\]

### 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 \( \epsilon \), 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:

\[
D_0 = \lim_{\epsilon \to 0} \frac{\log N_\epsilon}{\log(1/\epsilon)}
\]

### 5.1.3. Higuchi’s dimension

The Higuchi method [48] calculates fractal dimension from the estimated length of a signal. From an original signal \( x(1), x(2), \ldots, x(N) \) of length \( N \) a new signal, \( X_k^{\epsilon} \), is constructed: \( X_k^{\epsilon} = [x(m), x(m+k), x(m+2k), \ldots, x(m+[(N-m)/k])] \) (\( m = 1, 2, \ldots, k \)), 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 \( X_k^{\epsilon} \) 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 \( L_m(k) \):

\[
\langle L(k) \rangle = \frac{\sum_{m=1}^{k} L_m(k)}{k}
\]

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 \( \Delta t \) is the time increment. The variance \( \sigma^2 \) is related to the time increments \( \Delta t \) of signal \( x(t) \) according to the power law [49].

\[
\text{Var}[\Delta x(t, \Delta t)] = \propto |\Delta t|^{2H},
\]

where
\( \Delta x(t_\nu, \Delta t) = x(t_\nu + \Delta t) - x(t_\nu) \) and \( H \) is the Hurst exponent computed from a log-log plot using

\[
H = \lim_{\Delta t \to 0} \frac{1}{2} \log \frac{\text{Var}[\Delta x(t_\nu, \Delta t)]}{\log(\Delta t)}
\]

Finally, the variance dimension is defined as \( D_r = E + 1 - H \), 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 \( u_n(i) \). The number of vectors \( u_n(i) \) and \( u_n(j) \), close to each other, in an Euclidean sense \( \| u_n(i) - u_n(j) \| \leq r \), is expressed by the number \( n(r) \). This number is used to calculate the probability of vectors being close according to \( P(r) = n(r)/(N - m + 1) \). Let’s define the function \( \Phi^m(r) = 1/(N - m + 1) \sum_{i=1}^{N-m+1} \ln P^m(r) \). Consequently the ApEn can be defined as

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

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 \( N \geq 1000 \).

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:

\[
\text{SampEn}(m, r) = \lim_{N \to \infty} - \ln \frac{C^m r}{C^m 0}.
\]

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, \( x(i+1) \leq x(i) \), 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 ($= \text{positive predictive value}$) $PR = TP/(TP + FP)$ and recall ($= \text{sensitivity}$) $RE = TP/(TP + FN)$:

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

Parameters $TP$, $FP$, and $FN$ stand for true positive, false positive and false negative, respectively. The parameter $\beta$ 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 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 $\tau = 3.9 \pm 2.5s$. 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 Figure 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] (20x20 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 Figure 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.

$$PR = TP/(TP + FP) \quad \text{and} \quad RE = TP/(TP + FN).$$

$$F_\beta = \frac{(1 + \beta^2)(PR \cdot RE)}{\beta^2 \cdot PR + RE}.$$
RR intervals from external (ultrasound) records were corrupted adult HRV \[64\]. There are two possible explanations. First, the we can use correlation dimension \(D\) ties especially when using nonlinear methods. As an example 9. Discussion

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

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

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.

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, 3, and 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 process 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 process. Nevertheless, it is evident that proper processing of the FHR will remain an important part of any future decision support system.

Acknowledgements

This work was supported by the research programs No. NTI11124-6/2010 of the Ministry of Health Care and by the research program MSM 6840770012 Transdisciplinary Research in the Area of Biomedical Engineering II of the CTU in Prague, sponsored by the Ministry of Education, Youth and Sports of Czech Republic.

References


[27] G. Georgoulas, D. Gavrilis, I. G. Tsoulis, C. D. Stylios, J. Bernardes, P. P. Groumpos, Novel approach for fetal heart rate classification intro-