Automatic evaluation of FHR recordings from CTU-UHB CTG database

Jiří Spilka[1] George Georgoulas[2] Petros Karvelis[2] Vangelis P. Oikonomou[2]
Václav Chudáček[1] Chrysostomos Stylios[2] Lenka Lhotská[1] Petr Janků[3]

[1] Dept. of Cybernetics, Czech Technical University in Prague, Czech Republic
[2] Dept. of Informatics and Communications Technology, TEI of Epirus, Arta, Greece
[3] Dept. of Gynecology and Obstetrics, Teaching Hospital of Masaryk University in Brno, Czech Republic
spilka.jiri@fel.cvut.cz

Abstract. Fetal heart rate (FHR) provides information about fetal well-being during labor. The FHR is usually the sole direct information channel from the fetus – undergoing the stress of labor – to the clinician who tries to detect possible ongoing hypoxia. For this paper, new CTU-UHB CTG database was used to compute more than 50 features. Features came from different domains ranging from classical morphological features based on FIGO guidelines to frequency-domain and non-linear features. Features were selected using the RELIEF (RELevance In Estimating Features) technique, and classified after applying Synthetic Minority Oversampling Technique (SMOTE) to the pathological class of the data. Nearest mean classifier with adaboost was used to obtain the final results. In results section besides the direct outcome of classification the top ten ranked features are presented.

Keywords: fetal heart rate, intrapartum, feature selection, classification

1 Introduction

Electronic fetal monitoring (EFM) is used for fetal surveillance during pregnancy and, more importantly, during delivery. The EFM most commonly refers to cardiotocography (CTG) that is a measurement of fetal heart rate (FHR) and uterine contractions (UC). Since its introduction the CTG has served as the main information channel providing obstetricians with insight into fetal well-being. CTG monitoring still plays a role of the most prevalent method in use for monitoring of antepartum as well as intrapartum fetal well-being. The goal of fetal monitoring is to prevent fetus of potential adverse outcomes and provide an information about his/hers well-being. The main advantage of CTG, when compared to previously used auscultation technique, lies in its ability of continuous fetal surveillance though, this advantage is claimed to be insignificant in preventing adverse outcomes (with exception of neonatal seizures) as described in meta-analysis of several clinical trials [1]. The other main controversies of CTG include: increased rate of cesarean sections [1] and high intra- and inter-observer variability [2, 3].

Nowadays CTG remains the most prevalent method for intrapartum fetal surveillance [2, 4], often supported by ST-analysis (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 [5, 6] but its use is not always possible or feasible since it requires invasive measurement. Moreover, in order to use ST-analysis the correct interpretation of CTG is still required.

The interpretation of CTG is based on FIGO guidelines [7] introduced in 1986, or their newer international alternatives [8]. The main goal of guidelines is to assure lowering of the number of asphyxiated neonates while keeping the number of unnecessary cesarean sections (due to false alarms) at possible minimum. Additional goal of the guidelines was to lower the high inter and intra-observer variability. Despite the efforts made, the variability of clinicians evaluation of CTG still persists [9]. Three possible ways to lower it were discussed. e.g. [10] i) by extensive training, ii) using the most experienced clinician as an oracle, iii) and/or by computerized system supporting clinicians with the decision process.

The attempts of computerized CTG interpretation are almost as old as the FIGO guideline themselves. Beginning with work of [11] the automatic analysis of CTG was aligned with clinical guidelines [12]. Beyond the morphological features used in the guidelines, new features were introduced for FHR analysis. These were mostly based on the research in the adult heart rate variability [13]. The statistical description (time domain) of CTG tracings was employed in [14] and in [15]. The spectrum of FHR either in antepartum or intrapartum period offered insight to fetal physiology, and the short review [16] described recent development in this area. The joint time-frequency analysis of FHR in the form of wavelet analysis was employed in [17]. Nonlinear methods are widely used for FHR analysis [18, 19] and in our recent work we showed their usefulness in this field [20]. Different approaches were used for classification of FHR into different categories either based on pH levels, base deficit, or other clinical parameters. These approaches includes: Support Vector Machines (SVMs) [17, 21, 20], artificial neural networks (ANNs) [22, 23], or a hybrid approach utilizing grammatical evolution [24].

The contributions of the paper are twofold: First, from the CTG point of view, the used database will be open access at the time of publication, This is one of the largest databases used for automatic evaluation of the CTG. Second, we provide a promising approach for the automatic classification of CTG using the umbilical pH value as a gold standard. The results could serve as a base methodology for a new algorithm development on clinically sound data. An overview of the procedure is shown in Fig. 1.

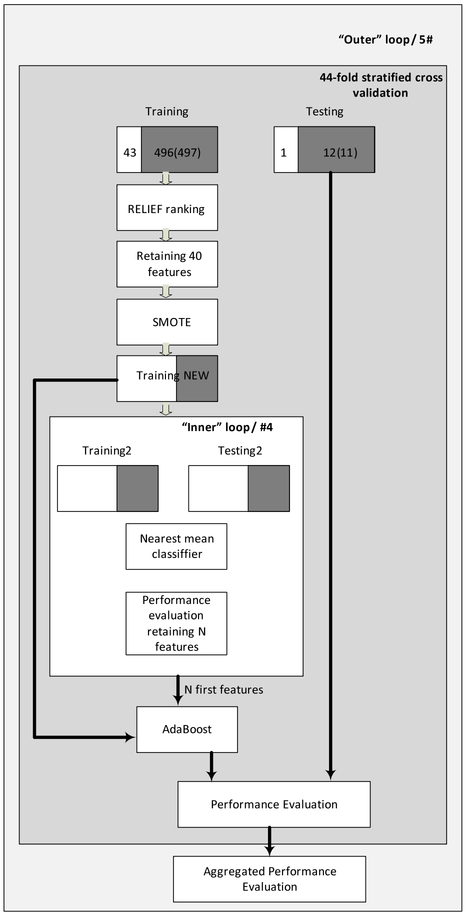


Fig. 1. An overview of the procedure.

2 Data Used

The database of 552 records is a subset of 9164 intrapartum CTG recordings that were acquired between years 2009 and 2012 at the obstetrics ward of the University Hospital in Brno, Czech Republic. The CTG signals were carefully selected with clinical as well as technical considerations in mind. The main parameters and their distributions are presented in Table 1. We have decided to select recordings that ended as close as possible to the birth and that had in the last 90 minutes of labor at least 40 minutes of usable signal. Additionally since CTG signal at II. stage of labor is very difficult to assess [25], we have included to the database only those recordings which had II. stage at maximum 30 minutes-long. The CTGs were recorded using STAN S31 (Neoventa Medical, Mölndal, Sweden) and Avalon FM40 and FM50 (Philips Healthcare, Andover, MA). The acqusition technique was either by scalp electrode (FECG 102 records), ultrasound probe (412 records), or combination of both (35 records). For three records the information was not available. All recordings were sampled at 4Hz by a recording device. The majority of babies were delivered vaginally (506) and rest using caesarean section (46). The more detailed description is provided in [26].

From the 552 recordings, 44 of them had pH value lower or equal to 7.05 as this is most commonly used value for distinction between pathological and normal outcome in the literature cf. e.g. [5]. Other thresholds were used in research work, for overview see e.g. [26]. This threshold value was selected for the formation of two classes. In this study we cast the assessment of fetus wellbeing as a classification problem.

Table 1. Overview of main parameters of the used CTU-UHB cardiotocography database

	Mean	Min.	Max.	Comment
Maternal age [years]	29,8	18	46	Over 36y: 40.
Parity	0,43	0	7	
Gravidity	1,43	1	11	
Gesta. age [weeks]	40	37	43	Over 42 weeks: 2
pH	7,23	6,85	7,47	Pat.: 48; Abnor.: 64
BDecf [mmol/l]	4,6	-3,4	26,11	Pat.: 25; Abnor.: 68
Apgar 5min	9,06	4	10	AS5 $<$ 7: 50
Neonate's weight [g]	3408	1970	4750	Small: 17; Large: 44
Neonate's sex [F/M]	259 / 293			

3 Signal Processing and Feature Extraction

3.1 Signal Preprocessing

The FHR was measured either externally using Doppler ultrasound (US) or internally by a scalp electrode (DECG); in special cases the combination of methods was used, i.e. beginning recording with US measurement and ending with DECG measurement. FHR recorded externally has lower signal to noise ratio than that recorded internally. The artifacts could be caused by mother/fetal movement, displacement of ultrasound probe, or simply by mis-detection of fetal heart beat by the recording device. We employed a simple artifacts rejection scheme: let $x(i)$ be a FHR signal in beats per minute (bpm), where N is number of samples and $i = 1, 2, \ldots, N$, whenever $x(i) \leq 50$ or $x(i) \geq 210$ we interpolated $x(i)$ using cubic Hermite spline interpolation. We used interpolation implemented in MATLAB®. We interpolated artifacts or missing data when the length of missing signal was equal or less than 15 seconds – the value based on FIGO guidelines and our experiments. When computing features we skipped the long gaps (> 15 seconds). An example of the result of artifacts removal is presented in Fig. 2.

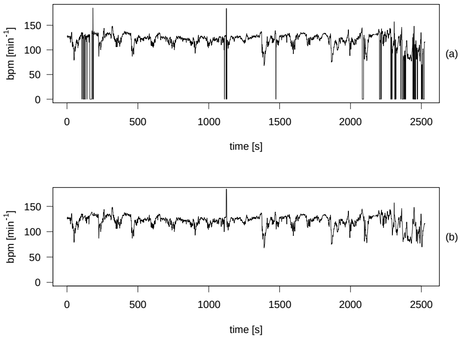


Fig. 2. Artefacts rejection. (a) Raw signal with artefacts, (b) signal after artefacts rejection.

3.2 Feature Extraction

As mentioned above the FIGO guidelines features were essential for the development of any system for automatic classification. Beyond that, other features, originating from different domains, were examined and used for classification. In this section we briefly describe the features, the description should serve as a context necessary to reproduce the analysis. We refer the interested reader to the referenced papers or to our previous works [20, 27].

Morphological features (used in clinical settings) Morphological features proposed in the FIGO guidelines represents macroscopic – "visible" – properties of the FHR. The morphological features were as follows: **mean of FHR baseline**, where the baseline is the mean level of fetal heart rate where acceleration and deceleration are absent; **number of accelerations**, where acceleration is a transient increase in heart rate above the baseline by 15 bpm or more, lasting 15 seconds or more; **number of decelerations**, where deceleration is a transient episode of slowing fetal heart rate below the baseline level by more than 15 bpm and lasting 10 seconds or more.

Short/long term variability The short term variability (STV) is the only feature sometimes computed automatically in clinical settings. The computation of STV depends whether FHR is recorder internally or externally. For the internal recording real beat-to-beat variability could be estimated while for external monitoring there is no real beat-to-beat (BB) variability because of intrinsic smoothing due to the correlation based technique. Instead epoch-to-epoch (EE) variation is used when the FHR is averaged over short period of time (2.5-3.75

sec.). Recall that $x(i)$ is the i-th FHR sample in beat per minute (bpm), let $T(i)$ a FHR sample in milliseconds $i = 1, 2, \ldots, N$, where N is the length of FHR. As noted in [28] the STV computed using $x(i)$ and $T(i)$ is not always the same because of dependence on the value of FHR mean utilized in some variability computation. The STV is estimated for signals of length 60 sec.; for longer signals the 60 sec. estimations are averaged. There exist several methods for computing STV and LTV, a comparison could be found in [28]. Here we present only a short list: **STVavg** estimated as the average of successive beat differences: $\mathrm{STV} = \frac{1}{N}\sum_{i=1}^{N-1} |T(i+1) - T(i)| \quad [\mathrm{ms}]$, **STV-DeHann** estimated as the inter quartile range of angular differences between successive $T(i)$s [29], **SDNN** [13], **STV-Yeh** [30], and **Sonicaid 8000** [31]. Long term variability (LTV) features were computed over 60 seconds and there was no need of averaging the FHR in 60 seconds. For FHR signals longer than 60 sec. estimations of LTV were averaged over each 60 sec. **LTV-DeHaan** [29] and the **Delta value** [14]. Many of the above mentioned features have been used in cases of antepartum signal evaluation and the effectiveness of many of them depends on their performance in the presence of accelerations and decelerations.

Frequency domain features Various spectral methods have been used for the analysis of adult heart rate [13]. In the case of FHR analysis, no standardized use of frequency bands exists. Therefore we used two slightly different partitionings of the frequency bands as was previously used in our work [21]. First we divided the frequency range into 3 bands [13] and calculated the energy of the signal in each one of them: **Very Low Frequency** (VLF); **Low Frequency** (LF) referred to as Mayer waves and **High Frequency** (HF) corresponding to fetal movement. Additionally the **ratio of energies** in the bands: ratio_LF_HF = $\frac{LF}{HF}$ was computed. It is a standard measure in adults and expresses the balance of behavior of the two autonomic nervous system branches. The alternative frequency partitioning followed suggestions of [32]. They proposed the following 4 bands: **Very Low Frequency** (VLF); **Low Frequency** (LF) correlated with neural sympathetic activity; **Movement Frequency** (MF), related to fetal movements and maternal breathing; **High Frequency** (HF), marking the presence of fetal breathing. Similarly to the previous 3-band division the following ratio of energies was computed: ratio_LF_MFHF = $\frac{LF}{MF+HF}$. This ratio is supposed to quantify the autonomic balance control mechanism (in accordance with the LF/HF ratio normally calculated in adults). The spectrum of FHR was estimated using the fast Fourier transform.

Nonlinear features Almost all nonlinear methods used for FHR analysis have their roots in adult HRV research. For nonlinear features we detrened FHR by the estimated baseline and also normalized the signal to have zero mean and unit variance. The **Poincaré plot** is a basic nonlinear feature commonly used in HRV domain [13]. The plot is a geometric representation of HRV where each RR interval is plotted as a function of the previous one. In this work we estimated waveform fractal dimension by several methods. These were: **box-counting dimension**, which expresses the relationship between the number of boxes that contain part of a signal and the size of the boxes; **the Higuchi method (FD_Hig)** [33], where the curve length $\langle L(k)\rangle$ is computed for different steps k and it is related to the fractal dimension by an exponential formula; the **variance fractal dimension (FD_Var)** that is based on properties of fractional Brownian motion. The variance $sigma^2$ is related to the time increments Δt of a signal $X(t)$ according to the power law [34]; an estimate of the fractal dimension proposed by **Sevcik** [35]; Detrend Fluctuations Analysis (DFA) [36] for estimating the fractal dimension, D, via scaling exponent α, $D = 3 - \alpha$. For all methods, the fractal dimension was estimated as the slope of a fitted regression to log-log plot of, e.g. for Higuchi method $\langle L(k)\rangle$ versus k. Also we estimated two scaling regions corresponding to STV and LTV, respectively [33]. The separation (critical) time was 3s. In addition, in order to estimate both regions by one parameter, we also fitted the log-log plot with a second order polynomial which coefficients (first order and second order polynomial coefficient) correspond to the both STV and LTV. The **Approximate Entropy (ApEn)** is able to distinguish between low-dimensional deterministic systems, chaotic systems, stochastic and mixed systems [37]. ApEn(m,r) approximately equals the average of a natural logarithm of conditional probabilities that sequences of length m are close to each other, within a tolerance r, even if a new point is added. A slightly modified estimation of approximate entropy was proposed by [38] and resulted in **Sample Entropy (SampEn)**. This estimation overcame the shortcomings of the ApEn mainly because the self-matches were excluded. The used parameters for ApEn and SampEn estimation are: tolerance $r = (0.15; 0.2) \cdot SD$ (SD stands for standard deviation) and the embedding dimension $m = 2$ [39] The last of the nonlinear features was the **Lempel Ziv Complexity (LZC)** [40]. This method examines reoccurring patterns contained in the time series irrespective of time. A periodic signal has the same reoccurring patterns and low complexity while in random signals individual patterns are rarely repeated and signal complexity is high.

3.3 Feature Selection

Usually in most pattern recognition applications the feature extraction stage is followed by a feature selection stage [41] which reduces the input dimensionality, because in real world applications we tend to extract more features than necessary in an effort to include all possible information. However, sometimes some of the extracted features can be correlated, hence redundant information is likely to be included or sometimes some features are irrelevant to the application at hand and may negatively affect the performance of the classifier. The term "performance" refers to the training time required during the construction of the classification model or, which is the more serious side-effect, the discriminative capability of the classifier. In feature selection, a search problem of finding a subset of l features from a given set of d features, $l < d$ has to be solved in order to optimize a specific evaluation measure, i.e the performance of the classifier. There are a number of approaches that try to tackle this problem which can roughly be divided into three categories: filters, wrappers and embedded methods [42]. The filter approach ranks features based on a performance evaluation metric calculated directly from the data; the wrapper approach employs a predictive model and uses its output to determine the quality of the selected features and the embedded approach integrates the selection of features in model building. In this work a hybrid approach combining a filter and a wrapper approach was combined. More specifically RELevance In Estimating Features (RELIEF) was employed to rank the features and then based on the ranking the number of retained features was determined by directly estimating their performance using a predictive model. In the rest of the section we briefly present RELIEF whereas the wrapped stage is explained in more detail in Section 4.

RELIEF is a popular feature selection algorithm based on a weight vector over all features which is updated according to the sample points presented (the higher the weight the better the feature). The algorithm for a binary classification problem can be summarized as follows

Algorithm 1: RELIEF algorithm

```
Input: a data set $D =< \mathbf{x_1}, y_1, \ldots, \mathbf{x_M}, y_M >$, with $\mathbf{x_i} \in \mathbb{R}^N$ and $y_i \in \{-1, 1\}$
        for $i = 1, \ldots, M$
a relevancy cut-off (threshold) $\tau$
a number of iteration $T$
begin
    i) initialize the weight vector to zero $\mathbf{w} = (0, 0, \ldots, 0)$
    ii) for $\underline{t \in T}$ do
        pick at random an example $\mathbf{x}$
        for $\underline{i \in N}$ do
            update the elements of the weight vector
            $w_i = w_i + (x_i - nearmiss(\mathbf{x})_i)^2 - (x_i - nearhit(\mathbf{x})_i)^2$
            where $nearmiss(\mathbf{x})$ and $nearhit(\mathbf{x})$ denote the nearest point to $\mathbf{x}$ in
            $D$ that belong to the other and the same class, respectively.
        end
        iii) select the feature set whose members exceed the given relevancy
        cut-off (threshold) $\tau$, $S = \{i | w_i > \tau\}$
    end
end
```

In our case the step *iii*) was not involved. Instead we selected the highest 40 out of the total 54 features and then we employed a wrapper approach using the simplest form of search procedure, the "Best Individual" [43], in order to select the number of retained features. In other words after using RELIEF to rank the features we tested 40 different subsets starting from a subset containing the feature with the highest rank and we continued adding one feature at a time (the second best, the third best etc.) and we estimated their classification performance. The subset with the highest performance was determining the number of features involved in the estimation performance phase as it is will be presented in more detail in the next Section 4.

4 Classification Procedure

As it was pointed out in section 2, one class, the abnormal one, is heavily undersampled in comparison to the normal one. This creates an extra challenge to the already difficult task of fetus well-being diagnosis. The class imbalance is a fundamental problem, arising when pattern recognition methods are dealing with real life problems, and many approaches have been proposed to overcome this situation [44]. In order to compensate for this imbalance we employed a popular technique which operates on the minority class creating artificial data, the Synthetic Minority Oversampling TEchnique (SMOTE). SMOTE is based on real data belonging to the minority class and it operates in the feature space rather than the data space [45]. The algorithm for each instance (in feature space) of the minority class introduces a synthetic example along any/all of the lines joining that particular instance with its k nearest neighbors that belong to the minority class. Usually after SMOTE the training set has approximately equal numbers of the 2 classes. However in this study our preliminary results suggested that more synthetic data from the minority class were needed. Therefore we selected to oversample the minority class by a factor of 18 using 27 (k=27) neighbors without trying to further optimize/tune the parameter settings of SMOTE. For testing the classification performance by making use of as many of the abnormal instances as possible we applied a 44 fold (stratified) cross validation procedure with each fold containing 1 abnormal instance and 12 or 11 normal instances. Therefore each time 43 abnormal instances were used for training and 496 or 497 normal instances and 1 abnormal instance and 12 (11) normal instances were saved for testing. During each fold we applied SMOTE to the abnormal instances, with the aforementioned parameters, while the normal instances remained intact. After the application of SMOTE an ”inner” loop involved for the selection of the ”optimal” number of features. During every fold RELIEF used all the training data (not the synthetic ones) to rank the features and then an inner loop was executed 4 times during which the data was randomly divided into training and testing (70/30) and a classifier was tested using 1 to 40 features (starting with the best feature and adding one feature at a time based on its ranking). Based on the average classification accuracy over these four repetitions the ”optimal” number of features was selected. After selecting the number of retained features, the whole training set (with the inclusion of the data coming from the SMOTE stage) was used to train a classifier to be tested on the reserved testing set. In this work we employed the simplest member of the nearest prototype classifier family, the nearest mean prototype classifier, which assigns an instance to the class whose mean vector is closest to, during the inner loop procedure, and after that (after the selection of the number of features to retain) we employed the same classifier but within the adaboost framework in order to come up with a more powerful classification scheme. Adaboost which comes from adaptive boosting was first introduced by Freund and Schapire [46] is a general method for improving the performance of a week learner. It is the most well-known model guided instance selection for building ensemble classifiers. The basic steps of the algorithm are summarized as follows (following mainly the notation provided in [47]).

```
Algorithm 2: Adaboost algorithm
  Input: a data set $D =< \mathbf{x_1}, y_1, \ldots, \mathbf{x_M}, y_M >$, with $\mathbf{x_i} \in \mathbb{R}^N$ and $y_i \in \{-1, 1\}$
         for $i = 1, \ldots, M$
  $k_{max}$ – the maximum number of weak learners to be included in the ensemble
  $C$ – a weak learner
  begin
      i) initialize the weight vector $\mathbf{W}_1 = (1/M, 1/M, \ldots, 1/M)$
      ii) for $k = 1, \ldots, k_{max}$ do
          train weak learner $C_k$ sampling $D$ according to $\mathbf{W}_k$
          $e_k \leftarrow \sum_{i:C_k(x_i) \neq y_i} \mathbf{W}_k(i)$, where $C_k(\mathbf{x})$ is the output of the weak
          classifier for instance $\mathbf{x}$
          $\alpha_k \leftarrow \frac{1}{2} \ln\left(\frac{1-e_k}{e_k}\right)$
          $\mathbf{W}_{k+1}(i) \leftarrow \frac{\mathbf{W}_k(i)}{Z_k} \times \begin{cases} e^{-\alpha_k}, \text{ if } C_k(\mathbf{x_i}) = y_i \\ e^{-\alpha_k}, \text{ if } C_k(\mathbf{x_i}) \neq y_i \end{cases}$, $Z_k$ is normalizing constant
      end
      iii) classify any new instance $\mathbf{x}$ using $G(\mathbf{x}) = sign\left(\sum_{k=1}^{k_{max}} \alpha_k C_k(\mathbf{x})\right)$
  end
```

In this work 40 nearest mean classifiers were employed. Trying to avoid any bias regarding the selection of the normal instance in each fold we repeated the procedure 5 times each time randomly reshuffling the normal instances creating an "outer" loop. By the outer/inner scheme we decouple the parameter selection stage from the estimation of the performance in an attempt to avoid getting optimistic results. The overall procedure is depicted in Fig. 1. The results of the 5 times repetition of the stratified 44-fold cross-validation procedure are summarized in the aggregated/cumulative confusion matrix, see Tab. 2.

As it can be observed with the specific setting we managed to have a balanced performance for both the normal and abnormal case. Regarding the feature selection process, Fig. 3 shows the number of times each feature configuration (number of features) has been selected over the 5x44 trials. As it can be seen usually a number between 10 and 20 was the most frequent configuration.

Regarding the ranking of the features by the RELIEF algorithm, Fig. 4a depicts the average ranking of the features (lower values are better) whereas Fig. 4b shows the number of times each individual feature was ranked.

Table 3 summarizes the top 10 features in terms of their average rank and Tab. 4 summarizes the top 10 features in terms of occurrences within the top 20 spot list.

Table 2. Cumulative confusion matrix of the proposed approach.

		Predicted	
		Abnormal	Normal
Actual	Abnormal	141	79
	Normal	884	1656

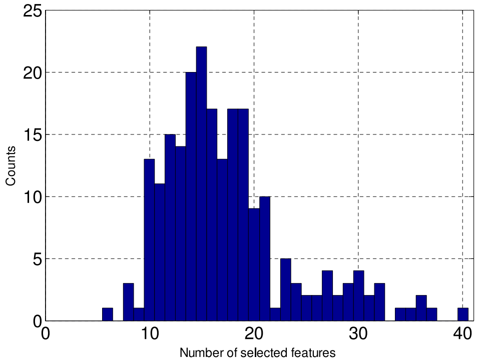
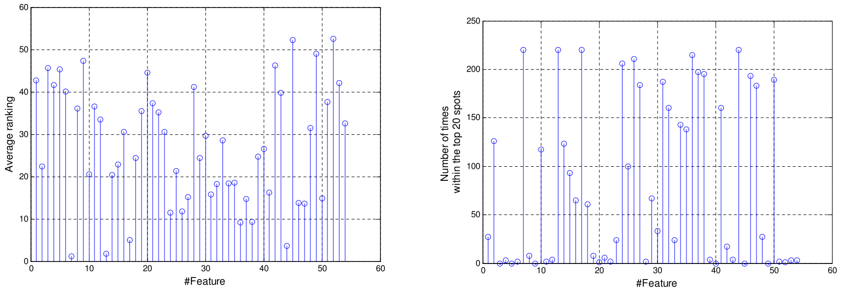


Fig. 3. Histogram of the number of selected features, through the "inner" loop procedure.



(a)

(b)

Fig. 4. Features ranking. a) Average ranking of individual features, b) Number of occurrences of individual features within the top-20 ranked list

Table 3. The top 10 features selected by the RELIEF algorithm.

Feature	7	13	44	17	36	38	24	26	47	46
Average rank	1.21	1.84	3.60	4.96	9.17	9.27	11.40	11.81	13.55	13.80

The numbers in the Table. 3 stands for the following features: **7** – STV-DeHann, **13** – meanBaseline (mean of FHR baseline), **44** – energy04_LF (low frequency energy for four bands division), **17** – lzc (Lempel Ziv Complexity),

Table 4. The top 10 features in terms of occurrences within the top 20 list.

Feature	7	13	44	17	36	26	24	37	38	46
# occurrences	220	220	220	220	215	211	206	197	195	193

36: DFA_p1 (detrend fluctuation analysis estimated second order polynomial coefficient), **26** – BoxCount_p1 (second order polynomial coefficient estimate by box counting method), **24** – BoxCount_Ds (box counting fractal dimension on short scale), **37** – DFA_p2 (detrend fluctuation analysis estimated first order polynomial coefficient), **38** – Sevcik fractal dimension, **46** – energy04_HF_LF (high frequency energy for four bands division).

5 Discussion and Conclusion

In this work we have used a broad range of features originating from different domains (time, frequency, state-space) for classification of CTG records into normal and abnormal classes. We used a database of CTG records, which is one of the largest database in the research field of CTG signal processing and classification. We implemented a hybrid filter-wrapper approach for feature selection were roughly 25% of features was filtered out using RELIEF algorithm and the rest were coupled with a nearest mean prototype classifier for further redcing the dimensionality of the input space. Our results indicate that the probably the selection was too conservative and further reduction might be possibly useful. The best selected features come from various domains, with the nonlinear features being the most prevalent. This corresponds to our previous results [20, 27], even though that previous study was performed on different database with smaller number of instances. Even though the linear correlation between features and pH is not high at all as shown in [48] and confirmed by our own tests, in our case, we managed to have a relatively high classification performance (taking into account the low correlation of the features with the monitored parameter). Especially the four best ranked features 3 (STV-DeHann, meanBaseline, energy04_LF, and lzc) possess the most valuable information regarding discrimination between normal and abnormal cases.

As we mentioned the results can serve as a base for comparing more elaborate classification schemes involving differt feature selection schemes and different classifiers. In our future work we intend to use other ranking methods such as the minimal-redundancy-maximal-relevance (mRMR) criterion [49], which can cope better with the problem of redundant information and we also intend to try replace the wrapper approach and use a random forest (RF) [50] to act upon a reduced number of features since the results of this work suggest that around 20 features could be a reasonable set of features, reducing this way the computational burden of the wrapper approch by taking advantage of the relatively quick trainig time of the RF. Moreover other state of the art classifiers such as the SVMs and the deep belief neural networks will be tested and compared against the "base" results that were derived from this study.

Acknowledgments

This work was supported by the following research programs: research grant No.NT11124-6/2010 of the Ministry of Health Care of the Czech Republic, MSMT-11646/2012-36 of the Ministry of Education, Youth and Sports of the Czech Republic, CVUT grant SGS13/203/OHK3/3T/13, and by the joint research project "Intelligent System for Automatic CardioTocoGraphic Data Analysis and Evaluation using State of the Art Computational Intelligence Techniques" by the programme "Greece-Czech Joint Research and Technology projects 2011-2013".

References

1. Alfirevic, Z., Devane, D., Gyte, G.M.L.: Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. Cochrane Database Syst Rev **3**(3) (2006) CD006066
2. Bernardes, J., Costa-Pereira, A., de Campos, D.A., van Geijn, H.P., Pereira-Leite, L.: Evaluation of interobserver agreement of cardiotocograms. Int J Gynaecol Obstet **57**(1) (Apr 1997) 33–37
3. Blix, E., Sviggum, O., Koss, K.S., Oian, P.: Inter-observer variation in assessment of 845 labour admission tests: comparison between midwives and obstetricians in the clinical setting and two experts. BJOG **110**(1) (Jan 2003) 1–5
4. Chen, H.Y., Chauhan, S.P., Ananth, C.V., Vintzileos, A.M., Abuhamad, A.Z.: Electronic fetal heart rate monitoring and its relationship to neonatal and infant mortality in the United States. Am J Obstet Gynecol **204**(6) (Jun 2011) 491.e1–491.10
5. Norén, H., Amer-Wåhlin, I., Hagberg, H., Herbst, A., Kjellmer, I., Maršál, K., Olofsson, P., Rosén, K.G.: Fetal electrocardiography in labor and neonatal outcome: data from the Swedish randomized controlled trial on intrapartum fetal monitoring. Am J Obstet Gynecol **188**(1) (Jan 2003) 183–192
6. Amer-Wåhlin, I., Maršál, K.: ST analysis of fetal electrocardiography in labor. Seminars in Fetal and Neonatal Medicine **16**(1) (2011) 29–35
7. FIGO: Guidelines for the Use of Fetal Monitoring. International Journal of Gynecology & Obstetrics **25** (1986) 159–167
8. ACOG: American College of Obstetricians and Gynecologists Practice Bulletin No. 106: Intrapartum fetal heart rate monitoring: nomenclature, interpretation, and general management principles. Obstet Gynecol **114**(1) (Jul 2009) 192–202
9. Blackwell, S.C., Grobman, W.A., Antoniewicz, L., Hutchinson, M., Gyamfi Bannerman, C.: Interobserver and intraobserver reliability of the NICHD 3-Tier Fetal Heart Rate Interpretation System. Am J Obstet Gynecol **205**(4) (Oct 2011) 378.e1–378.e5
10. de Campos, D.A., Ugwumadu, A., Banfield, P., Lynch, P., Amin, P., Horwell, D., Costa, A., Santos, C., Bernardes, J., Rosen, K.: A randomised clinical trial of intrapartum fetal monitoring with computer analysis and alerts versus previously available monitoring. BMC Pregnancy Childbirth **10** (2010) 71
11. Dawes, G.S., Visser, G.H., Goodman, J.D., Redman, C.W.: Numerical analysis of the human fetal heart rate: the quality of ultrasound records. Am J Obstet Gynecol **141**(1) (Sep 1981) 43–52

12. de Campos, D.A., Sousa, P., Costa, A., Bernardes, J.: Omniview-SisPorto 3.5 - A central fetal monitoring station with online alerts based on computerized cardiotocogram+ST event analysis. Journal of Perinatal Medicine **36**(3) (2008) 260–264
13. Task-Force: Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Eur Heart J **17**(3) (Mar 1996) 354–381
14. Magenes, G., Signorini, M.G., Arduini, D.: Classification of cardiotocographic records by neural networks. In: Proc. IEEE-INNS-ENNS International Joint Conference on Neural Networks IJCNN 2000. Volume 3. (2000) 637–641
15. Gonçalves, H., Rocha, A.P., de Campos, D.A., Bernardes, J.: Linear and nonlinear fetal heart rate analysis of normal and acidemic fetuses in the minutes preceding delivery. Med Biol Eng Comput **44**(10) (Oct 2006) 847–855
16. Van Laar, J., Porath, M., Peters, C., Oei, S.: Spectral analysis of fetal heart rate variability for fetal surveillance: Review of the literature. Acta Obstetricia et Gynecologica Scandinavica **87**(3) (2008) 300–306
17. Georgoulas, G., Stylios, C.D., Groumpos, P.P.: Feature Extraction and Classification of Fetal Heart Rate Using Wavelet Analysis and Support Vector Machines. International Journal on Artificial Intelligence Tools **15** (2005) 411–432
18. Ferrario, M., Signorini, M.G., Magenes, G., Cerutti, S.: Comparison of entropy-based regularity estimators: application to the fetal heart rate signal for the identification of fetal distress. IEEE Trans Biomed Eng **53**(1) (2006) 119–125
19. Gonçalves, H., Bernardes, J., Rocha, A.P., de Campos, D.A.: Linear and nonlinear analysis of heart rate patterns associated with fetal behavioral states in the antepartum period. Early Hum Dev **83**(9) (Sep 2007) 585–591
20. Spilka, J., Chudáček, V., Koucký, M., Lhotská, L., Huptych, M., Janků, P., Georgoulas, G., Stylios, C.: Using nonlinear features for fetal heart rate classification. Biomedical Signal Processing and Control **7**(4) (2012) 350–357
21. Georgoulas, G., Stylios, C.D., Groumpos, P.P.: Predicting the risk of metabolic acidosis for newborns based on fetal heart rate signal classification using support vector machines. IEEE Trans Biomed Eng **53**(5) (May 2006) 875–884
22. Czabanski, R., Jezewski, M., Wrobel, J., Jezewski, J., Horoba, K.: Predicting the risk of low-fetal birth weight from cardiotocographic signals using ANBLIR system with deterministic annealing and epsilon-insensitive learning. IEEE Trans Inf Technol Biomed **14**(4) (Jul 2010) 1062–1074
23. Georgieva, A., Payne, S.J., Moulden, M., Redman, C.W.G.: Artificial neural networks applied to fetal monitoring in labour. Neural Computing and Applications **22**(1) (2013) 85–93
24. Georgoulas, G., Gavrilis, D., Tsoulos, I.G., Stylios, C.D., Bernardes, J., Groumpos, P.P.: Novel approach for fetal heart rate classification introducing grammatical evolution. Biomedical Signal Processing and Control **2** (2007) 69–79
25. Sheiner, E., Hadar, A., Hallak, M., Katz, M., Mazor, M., Shoham-Vardi, I.: Clinical significance of fetal heart rate tracings during the second stage of labor. Obstet Gynecol **97**(5 Pt 1) (May 2001) 747–752
26. Chudáček, V., Spilka, J., Burša, M., Janků, P., Hruban, L., Huptych, M., Lhotská, L.: Open access intrapartum CTG database: Stepping stone towards generalization of technical findings on CTG signals. PLoS ONE **Manuscript submitted for publication** (2013)

27. Chudáček, V., Spilka, J., Lhotská, L., Janků, P., Koucký, M., Huptych, M., Burša, M.: Assessment of features for automatic CTG analysis based on expert annotation. Conf Proc IEEE Eng Med Biol Soc **2011** (2011) 6051–6054
28. Cesarelli, M., Romano, M., Bifulco, P.: Comparison of short term variability indexes in cardiotocographic foetal monitoring. Comput Biol Med **39**(2) (Feb 2009) 106–118
29. de Haan, J., van Bemmel, J., Versteeg, B., Veth, A., Stolte, L., Janssens, J., Eskes, T.: Quantitative evaluation of fetal heart rate patterns. I. Processing methods. European Journal of Obstetrics and Gynecology and Reproductive Biology **1**(3) (1971) 95–102 cited By (since 1996) 13.
30. Yeh, S.Y., Forsythe, A., Hon, E.H.: Quantification of fetal heart beat-to-beat interval differences. Obstet Gynecol **41**(3) (Mar 1973) 355–363
31. Pardey, J., Moulden, M., Redman, C.W.G.: A computer system for the numerical analysis of nonstress tests. Am J Obstet Gynecol **186**(5) (May 2002) 1095–1103
32. Signorini, M.G., Magenes, G., Cerutti, S., Arduini, D.: Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings. IEEE Trans Biomed Eng **50**(3) (Mar 2003) 365–374
33. Higuchi, T.: Approach to an irregular time series on the basis of the fractal theory. Phys. D **31**(2) (1988) 277–283
34. Kinsner, W.: 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)
35. Sevcik, C.: A Procedure to Estimate the Fractal Dimension of Waveforms. Complexity International **5** (1998) –
36. Peng, C.K., Havlin, S., Stanley, H.E., Goldberger, A.L.: Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos **5**(1) (1995) 82–87
37. Pincus, S.: Approximate entropy (ApEn) as a complexity measure. Chaos **5 (1)** (1995) 110–117
38. Richman, J.S., Moorman, J.R.: Physiological time-series analysis using approximate entropy and sample entropy. Am J Physiol Heart Circ Physiol **278**(6) (Jun 2000) H2039–H2049
39. Pincus, S.M., Viscarello, R.R.: Approximate entropy: a regularity measure for fetal heart rate analysis. Obstet Gynecol **79**(2) (Feb 1992) 249–255
40. Lempel, A., Ziv, J.: On the complexity of finite sequences. IEEE Transactions on Information Theory **IT-22 (1)** (1976) 75–81
41. Theodoridis, S., Koutroumbas, K.: Pattern recognition, 4th Edition (2009)
42. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A.: Feature extraction: foundations and applications. Volume 207. Springer (2006)
43. Webb, A.R.: Statistical pattern recognition. Wiley (2003)
44. Chawla, N.V., Japkowicz, N., Kotcz, A.: Editorial: special issue on learning from imbalanced data sets. ACM SIGKDD Explorations Newsletter **6**(1) (2004) 1–6
45. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research **16** (2002) 321–357
46. Freund, Y., Schapire, R.E.: Experiments with a New Boosting Algorithm. In: International Conference on Machine Learning. (1996) 148–156
47. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification. New York: John Wiley, Section **10** (2001) l

48. Fulcher, B., Georgieva, A., Redman, C., Jones, N.: Highly comparative fetal heart rate analysis. In: Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE. (28 2012-sept. 1 2012) 3135 –3138
49. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, IEEE Transactions on **27**(8) (2005) 1226–1238
50. Breiman, L.: Random Forests. Machine Learning **45**(1) (2001) 5–32