Novel approach for fetal heart rate classification introducing grammatical evolution

George Georgoulas [a], Dimitris Gavrilis [b], Ioannis G. Tsoulos [c],
Chrysostomos Stylios [d,*], João Bernardes [e], Peter P. Groumpos [f]

[a] *School of Electrical & Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA*
[b] *Electrical & Computer Engineering Department, University of Patras, Rion 26500, Greece*
[c] *Department of Computer Science , University of Ioannina, P.O. Box 1186, Ioannina 45110, Greece*
[d] *Technological Educational Institution of Epirus, Informatics and Telecommunicatios Technology, Kostakioi, 47100 Artas, Greece*
[e] *Departamento de Ginecologia e Obstetrícia, Faculdade de Medicina da Universidade do Porto, Portugal*
[f] *Laboratory for Automation and Robotics, Department of Electrical and Computer Engineering, University of Patras, Rion 26500, Greece*

Abstract

Fetal heart rate (FHR) variations reflect the level of oxygenation and blood pressure of the fetus. Electronic Fetal Monitoring (EFM), the continuous monitoring of the FHR, was introduced into clinical practice in the late 1960s and since then it has been considered as an indispensable tool for fetal surveillance. However, EFM evaluation and its merit is still an open field of controversy, mainly because it is not consistently reproducible and effective. In this work, we present a novel method based on grammatical evolution to discriminate acidemic from normal fetuses, utilizing features extracted from the FHR signal during the minutes immediately preceding delivery. The proposed method identifies linear and nonlinear correlations among the originally extracted features and creates/constructs a set of new ones, which, in turn, feed a nonlinear classifier. The classifier, which also uses a hybrid method for training, along with the constructed features was tested using a set of real data achieving an overall performance of 90% (specificity = sensitivity = 90%).
© 2007 Elsevier Ltd. All rights reserved.

Keywords: Fetal heart rate; Genetic algorithm; Grammatical evolution; Multilayer perceptron; Feature construction; Classification

1. Introduction

Antepartum and intrapartum fetal surveillance is widely based on monitoring, analysis and evaluation of fetal heart rate (FHR). FHR is acquired by cardiotocographs along with the uterine activity (UA) and both signals are printed on a single strip of paper termed cardiotocogram (CTG). FHR is measured in beats/minute and it is acquired either by scalp electrodes, after the rupture of the membranes, or by an external sensor attached on the mother's abdomen [1]. Fig. 1 shows a typical CTG. FHR signal alterations are used to evaluate the fetal health condition, so as to early diagnose fetal stress and distress. In the latter case,

the obstetrician has to intervene to prevent potential compromise and irreversible damage, such as cerebral palsy or death. Especially during the crucial period of labor, FHR monitoring is used as the main screening test of the fetal acid base balance [2].

FHR monitoring has been used during the last four decades as the main electronic fetal surveillance test. However, extensive studies have shown significant inter-observer and intra-observer variation in FHR analysis and interpretation [3–5]. These inconsistencies in interpretation and the increase of false positive diagnosis have created skepticism. On the other hand, the advances in pattern recognition methods, along with new signal processing techniques, have paved the way towards automated approaches. Therefore, many researchers proposed computer-based systems, in an attempt to monitor and evaluate the condition of the fetus in a reliable, reproducible and effective way [6–31].

The paths followed by researchers in the field are quite diverse. Mantel et al. [8,9], as well as Taylor et al. [11], used an

* Corresponding author. Tel.: +30 2681050330.
E-mail addresses: ggeorgoulas@mail.gatech.edu (G. Georgoulas),
gavrilis@upatras.gr (D. Gavrilis), itsoulos@cs.uoi.gr (I.G. Tsoulos),
stylios@teleinfom.teiep.gr (C. Stylios), jbernardes@mail.telepac.pt
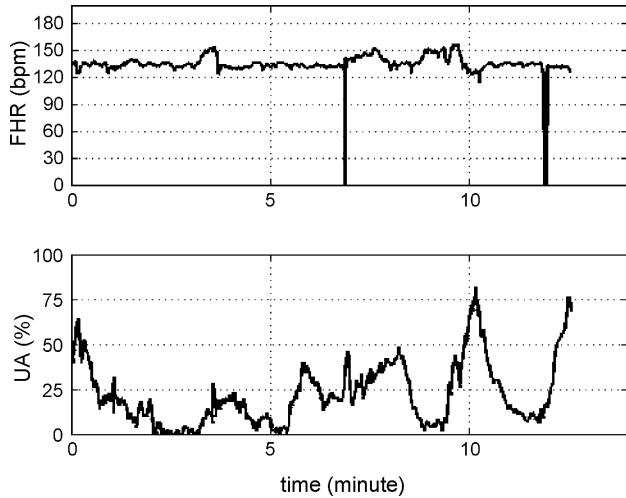(J. Bernardes), groumpos@ece.upatras.gr (P.P. Groumpos).

Fig. 1. A typical cardiotocogram.

iterative procedure to extract the main morphological features of FHR. Bernardes et al. [17–19] developed an automatic method for the analysis and recognition of CTG and subsequent extraction of morphological features. Magenes et al. [20,21] and Kol et al. [22] employed artificial neural networks for the interpretation of FHR recordings. Chung et al. [23] developed an algorithm to analyze and predict fetal acidosis. Salamalekis et al. [24], Struzik and Wijngaarden [25] and Georgoulas et al. [26] employed wavelets for the analysis of the FHR signal. Ifeachor et al. developed an expert system [27], which they subsequently upgraded into a fuzzy system [28], for the interpretation of CTG records. Alonso-Betanzos et al. merged FHR and contextual analysis of all pathological and physiological aspects involved in fetal monitoring. They developed at first an expert system called NST-EXPERT [29,30], which later evolved to create computer aided fetal evaluator (CAFE) [31].

Based on the belief that the FHR may convey much more information than what is usually interpreted by obstetricians [6], we investigate and propose an advanced methodology to analyze and interpret FHR. This method is used to early diagnose fetal acidemia. This innovative method utilizes a novel approach for feature construction based on the extracted features from the FHR signal. The constructed features are fed to a multilayer perceptron (MLP) nonlinear classifier [32,33], with very promising results.

The original FHR features are derived from three domains: the time domain, the frequency domain and the morphological domain. The last one utilizes medical definitions of "morphological" features, which have already been used with quite a success both in the antepartum [20,21] and intrapartum case [34,35].

The recently proposed approach of grammatical evolution is applied to construct new artificial features from the actual ones. Grammatical evolution [36] is an evolutionary methodology that, as in the case of genetic programming [37], can evolve complete programs. In our case, the evolved programs are mathematical expressions/functions of the originally extracted features. These constructed artificial features are then used to classify the FHR signal.

This paper is structured as follows: Section 2 gives a brief introduction to the grammatical evolution method. Section 3 presents the overall proposed procedure and the stages preceding the construction phase. Section 4 describes the implementation of the grammatical evolution for feature construction while Section 5 presents a brief analysis of the new features in terms of their "quality". Section 6 compares the experimental results for different implementations of the proposed scheme and, finally, Section VII concludes the paper and ideas and directions for future work are discussed.

2. Grammatical evolution

Grammatical evolution [36] uses an evolutionary algorithm and a Backus–Naur form (BNF) description [38] to create programs in an arbitrary language. In grammatical evolution, chromosomes are defined as a series of production rules of the appropriate BNF grammar. Each gene of the chromosome denotes a production rule from the BNF grammar. The chromosomes in grammatical evolution have variable size. The algorithm begins from the start symbol of the grammar and gradually creates the program string by replacing non-terminal symbols with the right hand of the selected production rule [36,39]. The selection of the appropriate rule is performed according to the following equation:

$$\text{Rule} = B \bmod RN \qquad (1)$$

where B is the specific gene and RN is the number of rules for the specific non-terminal symbol. This selection process is repeated and it stops when the end of the chromosome is reached. If the end of the chromosome is reached and no valid expression has been produced, the process continues again from the beginning of the chromosome (wrapping effect). The BNF grammar used in this work can be seen below in Table 1.

In our case, the BNF grammar uses mathematical functions and operators as non-terminal symbols and the original features $(x_1, x_2, \ldots, x_n)$ and the digits (0–9) as the terminal symbols. As a result, the expressions that can be produced from this grammar are mathematical expressions involving the operators $x_1, x_2, \ldots, x_n$ (which represent the original features and in our case are 19). The complete process of decoding a chromosome is summarized in Table 2.

The classical grammatical evolution algorithm differs from the one presented here. In this work, the size of chromosome is

Table 1
The BNF grammar involved for feature construction

Rule			Rule number
S	::=	\<expr\>	0
\<expr\>	::=	\<expr\> \<op\> \<expr\>	0
		|\<func\> (\<expr\>)	1
		|\<terminal\>	2
\<xlist\>	::=	$\|x_1\|x_2\|\ldots\|x_n$	$0\|1\|2\|\ldots\|n-1$
\<digitlist\>	::=	\<digit\> | \<digit\>\<digitlist\>	0|1
\<terminal\>	::=	\<xlist\> | \<digitlist\>.\<digitlist\>	0|1
\<op\>	::=	+|−|*|/	0|1|2|3
\<func\>	::=	sin | cos | exp | log | sqrt | abs	0|1|2|3|4|5
\<digit\>	::=	0|1|2|\ldots|9	0|\ldots|9

Table 2

Sequential creation of a feature decoding the original chromosome. Table 2 presents an example of the proposed method for feature construction. The original chromosome is: [9,8,6,4,16,10,17,23,8,14]. In the first iteration, starting from the start symbol S, the <expr> non-terminal is obtained. Because <expr> has three rules, 9 mod3 = 0 so <expr> is substituted by <expr> <op> <expr> which is rule number 0 (of the <expr> non-terminal). The next non-terminal (<expr>) is selected and the process is repeated until all non-terminals have been substituted with terminal symbols. The resulting expression is f(.) = x4 + cos(x14)

String	Chromosome	Operation
<expr>	9,8,6,4,16,10,17,23,8,14	9 mod 3 = 0
(<expr><op><expr>)	8,6,4,16,10,17,23,8,14	8 mod 3 = 2
(<terminal><op><expr>)	6,4,16,10,17,23,8,14	6 mod 2 = 0
(<xlist><op><expr>)	4,16,10,17,23,8,14	4 mod 19 = 4
(x4<op><expr>)	16,10,17,23,8,14	16 mod 4 = 0
(x4 + <expr>)	10,17,23,8,14	10 mod 3 = 1
(x4 + <func>(<expr>))	17,23,8,14	17 mod 4 = 1
(x4 + cos(<expr>))	23,8,14	23 mod 3 = 2
(x4 + cos(<terminal>))	8,14	8 mod 2 = 0
(x4 + cos(<xlist>))	14	14 mod 19 = 14
(x4 + cos(x14))		

fixed. The wrapping process is performed only once and if after the wrapping process still no valid expression has been extracted, the chromosome is rejected. These two modifications on the algorithm are made in order to increase speed and to avoid the creation of very large solutions (features). When the newly constructed features have been created, a new train and test set is attained from the original, according to the new features. A fitness value is assigned to each of the newly constructed features, which in our case is the accuracy of a supervised classifier. According to that fitness value, the newly constructed feature will be accepted or rejected in the successive generations.

We must point out that the selection of the fitness function is not unique. Many other measures reflecting the separability in the feature space could be involved (i.e. using scatter matrices [40]) or even functionals of measures of individual features. However, in this work we employ the "wrapper" approach where the construction of the features forms as a wrapper around the classification algorithm. Even though the term "wrapper" is usually involved in feature selection algorithms [41] we also use it here since the underlying philosophy is the same and has nothing to do with the "wrapping process" described in the previous paragraph. This proposed approach, even though it does not maximize separability or individual feature quality, it manages to achieve very high values of sensitivity and specificity, proving in a sense that "relevance" does not necessarily leads to high classification accuracy. By constructing a small number of new artificial features a reduction of the feature dimension is achieved, the curse of dimensionality is alleviated and the generalization performance is usually increased. More details about the proposed method can be found in [39].

3. Overall procedure

The overall proposed procedure constitutes an integrated approach, which takes as input the FHR signal and identifies the
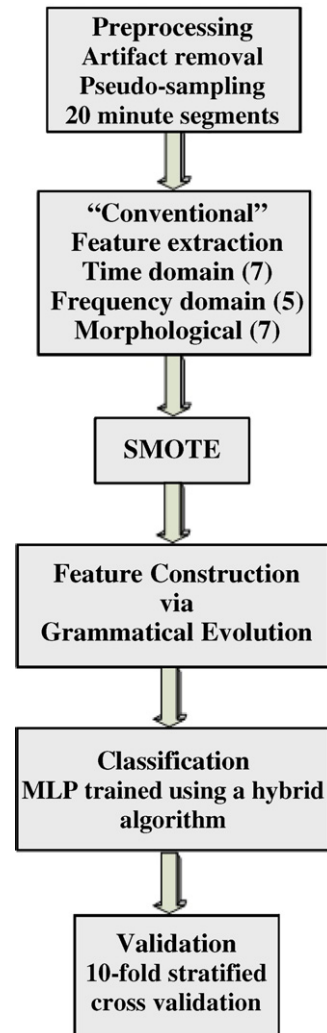

Fig. 2. The overall procedure.

health status of the fetus based on a prediction about its blood pH. Fig. 2 depicts the overall scheme, consisting of five stages and a final validation process.

3.1. Preprocessing stage

The preprocessing stage is divided into three sub-stages.

3.1.1. Artifact removal

The FHR signals undergo a removing of artifacts which are present in almost all recordings [24]. The artifact removal module, introduced in [17], firstly detects a stable FHR segment. Stable segment is defined as the segment where the difference between five adjacent samples is less than 10 beats/min. Whenever a difference between adjacent beats higher than 25 beats/min is found, a linear interpolation is applied between the first of those two segments and the first sample of a new stable FHR segment (Fig. 3).

3.1.2. Pseudo-sampling

Twenty-three recordings were acquired using a Toitu MT810B fetal monitor, which produces an irregularly sampled
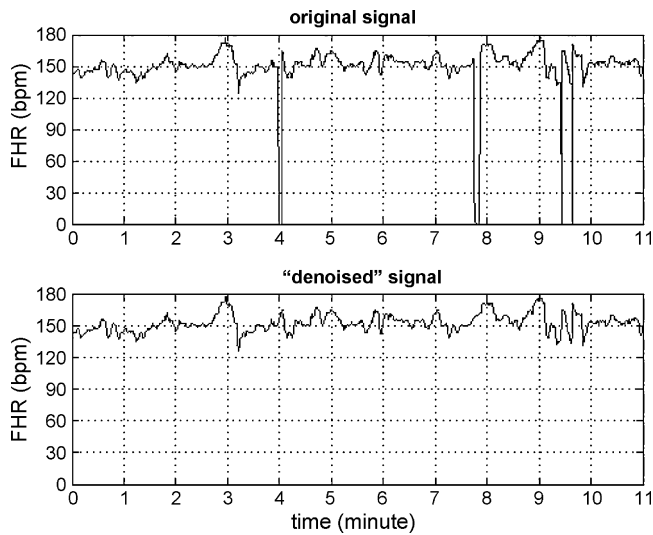
Fig. 3. FHR denoising. The top graph depicts the original FHR signal and the bottom graph the signal after removing spikes and filling the missing gaps.

time series. Those signals had to be transformed into regularly sampled ones through a pseudo-sampling procedure [34] before incorporating them into the larger available FHR data set. The second data set consisted of 137 recordings acquired by an HP 1350 fetal monitor operating at a sampling frequency of 4 Hz.

3.1.3. Segmentation

The FHR recordings had various lengths, ranging from 20 min to more than 1 h. In this work, we selected a time duration of 20 min as the time length of the segments to be analyzed. The 20 min segments were selected as close as possible to the end of the recordings (i.e. as close as possible to delivery) in order to avoid time bias.

3.2. Feature extraction stage

The selection of suitable features is probably the most important phase during the development of a classification system. As already pointed out, feature selection is more of an art than a science [42]. Therefore, sometimes it is worth testing a number of features before selecting an appropriate set. In this work, we go one step further and after the extraction of a number of features we construct new ones using the grammatical evolution phase. During this phase the features conveying the most valuable information will most probably be part of the expression for a newly constructed feature.

In an attempt to gather information from different domains, we selected a set of FHR features derived from (a) the time domain, (b) the frequency domain and (c) a set of morphological features. The last ones are closely related to the physician's approach for FHR interpretation which is directly associated with the "morphology" of the FHR signal. Subsequently, we will present the extracted features belonging to each one of the aforementioned three domains.

(1) Time domain features:
  x1. mean FHR;

x2. standard deviation of FHR;
  x3. delta value;
  x4. STV value;
  x5. II value;
  x6. LTI value;
  x7. delta total value.

A detailed description of each one of the above parameters can be found in [20,21,35].
(2) Frequency domain features
    In this research work, we adopted the partitioning of the frequency band proposed by Signorini et al. [6] because it is proven to give a more adequate characterization of the fetal condition [35]. Therefore, we partitioned the frequency range into the following four bands and we calculated the corresponding energies:
  x8. very low frequency (VLF) 0–0.03 Hz;
  x9. low frequency (LF) 0.03–0.15;
  x10. movement frequency (MF) 0.15–0.5 Hz;
  x11. high frequency (HF) 0.5–1 Hz;
  x12. as a fifth feature for this feature set, we used the ratio of energies: LF/(HF + MF).
(3) "Morphological" features
    Conventional interpretation of FHR is based upon certain morphological characteristics, according to the guidelines given in [43,44]. In this work, we examined the following set of morphological parameters:
  x13. baseline;
  x14. number of accelerations;
  x15. number of small accelerations;
  x16. number of mild decelerations;
  x17. number of prolonged decelerations;
  x18. number of severe decelerations;
  x19. the percentage of the time occupied by decelerations.

A detailed description of each one of the above parameters can be found in [19,35].
Therefore, we had a total of 19 initial features, which we used to construct the new artificial features through the process of grammatical evolution. However, before proceeding to this stage we had to take some action in order to alleviate the problem arising from the imbalanced nature of the available data.

3.3. Synthetic minority oversampling TEchnique (SMOTE) stage

In this research work, the data set consists of 160 FHR recordings. The recordings were collected in the context of the Research Project POSI/CPS/40153/2001, funded by Fundação para a Ciência e Tecnologia, Portugal. As already mentioned in Section 3, 137 recordings were acquired using an HP 1350 fetal monitor at a sampling frequency of 4 Hz, and 23 were acquired using a Toitu MT810B. In both cases, scalp electrodes were used for the acquisition, giving accurate recordings (greater accuracy compared to recordings acquired using ultrasound technology) [1,45]. One hundred and thirty FHR recordings
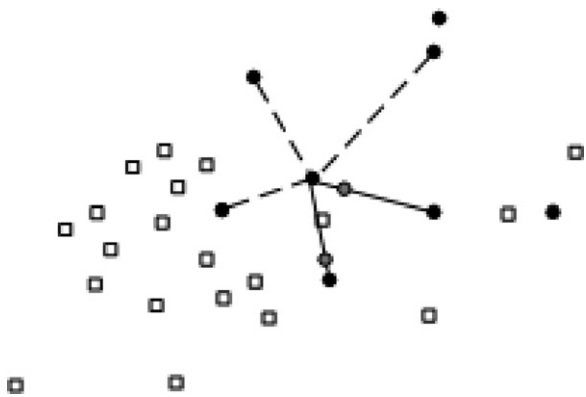
Fig. 4. Illustrative example of applying SMOTE. The minority class is marked with circles and the majority class with squares. The black circles indicate the original minority data while the grey circles indicate the produced ones.

belonged to normal fetuses, with umbilical artery pH > 7.20, and 30 recordings to fetuses with pH < 7.10, belonging to the "at risk" group.

As it is obvious, this segregation of data has an intrinsic imbalance and if no precaution is taken, the construction of the classifier may favor the class with the greatest number of instances, which in this case is the class with the normal subjects. 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 [46]. In this work, in order to alleviate this problem we adopted the procedure termed Synthetic Minority Oversampling TEchnique (SMOTE) which creates "synthetic" instances for the undersampled class. SMOTE is based on real data belonging to the minority class and it operates in the "feature space" rather than the "data space" [47].

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. A representative example for the implementation of SMOTE on some artificial data is depicted in Fig. 4. In this example, it is assumed that we want to have 200% oversampling and that the neighborhood of each minority consists of five samples.

For this particular work, we used eight nearest neighbors and we tested different amounts of oversampling percentages: 300, 400 and 500%.

4. Feature construction

After the SMOTE stage, we have more balanced feature sets which can be introduced to the grammatical evolution stage. The grammatical evolution procedure itself is divided into two phases: the construction and the evaluation phase.

4.1. Construction phase

A new artificial set of features is derived from the original feature set using a combination of grammatical evolution and an MLP neural network. The artificial features are constructed from the original ones using a nonlinear mapping. Using the newly constructed features, new training and testing sets are constructed based on the original sets.

The new features are then evaluated using the new train and test sets and an MLP [32,33] with one hidden layer and one neuron in the output layer. The MLP network is trained using Powell's Broyden–Fletcher–Golfard–Shanno (BFGS) variant algorithm [48]. During the feature construction process, the fitness function of each genome (i.e. the fitness of the corresponding constructed feature) is the classification accuracy of the neural network. In order to accelerate this process, only a small number of the BFGS procedure steps are performed. The genetic algorithm in the grammatical evolution procedure is evolved for 300 generations and each generation has 400 genomes. These parameters are empirically selected based on experiments utilizing 100–1000 generations and different population sizes. The chromosome length for each genome is fixed and it is set equal to 100.

However, as mentioned before, the original grammatical evolution algorithm uses variable length chromosomes. This approach was not adopted because variable length chromosomes produce very large expressions and it is much slower compared to the fixed length implementation. The mutation rate is set to 5% and the crossover rate to 95% (one point crossover). In each generation, the selection is performed using the tournament scheme with a tournament size of 10.

4.2. Evaluation phase

The derived features are then evaluated using a second neural network. The second neural network is also an MLP with one hidden layer. The network is trained for 1–10 hidden nodes in order to select the most efficient topology.

A hybrid two-step procedure is used to train the neural network, which is a combination of a genetic algorithm and a local optimizer. In the first step, a genetic algorithm is used to estimate the network's initial weights. The algorithm is used to find a good set of initial values and it is evolved for 50 generations with a population of 100 genomes. The crossover (one point crossover again), mutation and tournament size are 95%, 5% and 10, respectively. In the second step, the same BFGS variant is used during the construction phase and it is employed to refine the initial weights obtained by the genetic algorithm. The BFGS variant runs for a maximum of 2000 iterations. It is experimentally found that BFGS increases the overall efficiency. This hybrid approach incorporates the intrinsic characteristic of genetic algorithms to find the global minimum with the speed gained by the local optimizer [49]. In Fig. 5, the evolution of the fitness value for the best individual during the construction of one feature for the case of 300% oversampling is illustrated.

4.3. Validation

In order to validate the proposed method, we employed the 10-fold stratified cross validation method [50]. To be more
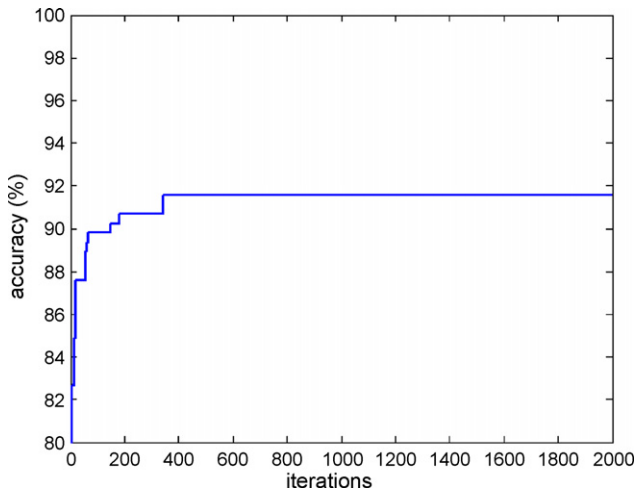
Fig. 5. Evolution of the fitness function (classification accuracy) during the BFGS cycle.

5. Results of the feature construction phase

As explained in Section 4, the feature construction is indispensably related to the training of the MLP and as a result the overall success of this approach is based on the synergy of these two components. This can further be highlighted when investigating the ''usefulness'' of individually constructed features. One common measure to evaluate the ''usefulness'' of an individual feature is based on the Fisher Discriminant Ratio (FDR) [40]:

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{2}$$

where the subscripts 1 and 2 refer to the mean and variance corresponding to the feature under investigation for the two classes ω_1, ω_2, respectively.

The larger the value of the FDR, the greater the separabiblity capabilities of the individual feature. In terms of this measure, the ''best'' feature of the original nineteen features is the II value (x5).

Another measure that can be used to assess the individual performance of a single feature is the area under the curve of the receiver operating characteristic curve (AUC) [40]. In terms of this measure the best original feature is the std (x2) which also has a relative high FDR value. The normalized histograms (an approximation of the class conditional probabilities) of the aforementioned features (x2, x5) are shown in Fig. 6

During the construction process, some artificial features had greater values both in terms of FDR and AUC, whereas some others had very low. For the case of the FDR, in one case the value was almost two times the value corresponding to the II feature (which also happens to have the maximum AUC). For that feature the normalized histogram as well as the ROC curve are depicted in Fig. 7.

We must point out that no consistent pattern can be found in the way those measures vary with increasing number of features and increasing synthetic examples. This, we speculate, happens for two main reasons. First of all the algorithm was not designed to maximize those quantities and, secondly, the use of the MLP
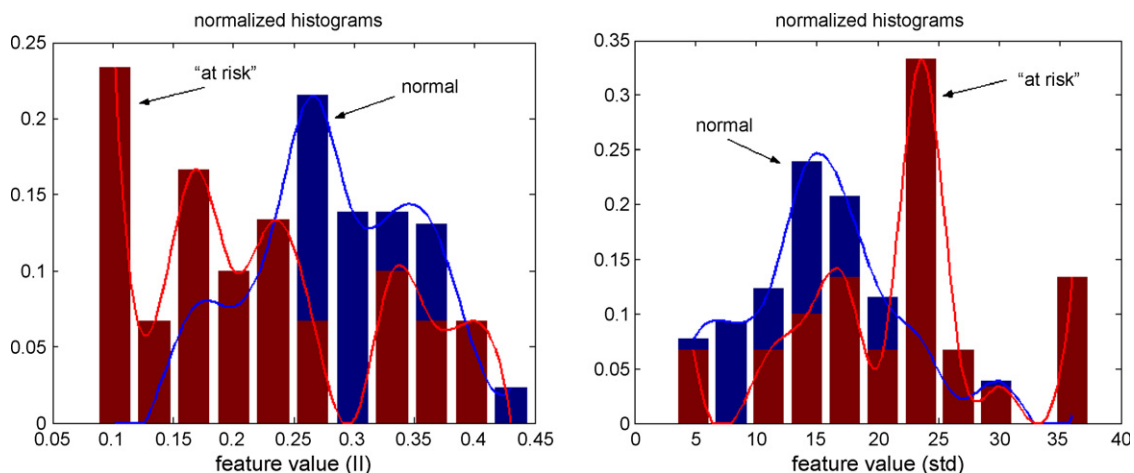
specific, we divided the 160 samples into 10 non-overlapping sets, each one containing 13 samples from the majority class and three samples from the minority class. Each time one set was excluded from the training phase and was used only for the estimation of the classification performance. The remaining nine sets (consisting of 107 examples from the normal group and 27 of the at risk group) were used for training. Before analyzing those 134 samples, we applied SMOTE to the 27 samples creating a number of synthetic examples depending on the amount of oversampling (i.e. 300% oversampling implies the creation of $27 \times 3 = 81$ ''new'' examples). Those examples plus the 27 original data comprised the set of the ''at risk'' cases that would be used for training. Due to the stochastic nature of the grammatical evolution method, we repeated this procedure 10 times and we took the mean value over these 10 iterations. It must be mentioned that we paid much attention so as to avoid using the same data set both for tuning and estimating the performance of the proposed classifier [51]. Thus, we employed the cross validation procedure within the training set so as to find the best architecture (features + MLP structure) that we employed afterwards using the whole training set.

Fig. 6. Normalized histograms for two of the more ''relevant'' features of the original feature set (the overlaid curves are only used for visualization purposes).
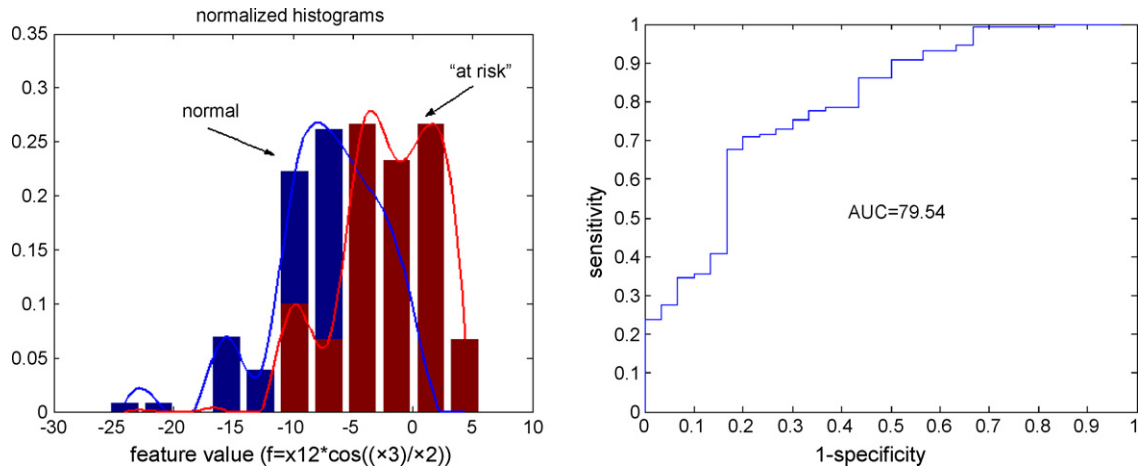
Fig. 7. Normalized histograms for constructed feature with the maximum FDR and AUC.

with a hidden layer acts as a second level of feature extraction making the analysis more complicated. In other words, the MLP itself performs a second level mapping through its hidden layers, which cannot be depicted in the feature distribution of the original feature space.

It must be mentioned that the feature construction algorithm showed preference to some initial features, which in terms of the FDR and the AUC, did not seem to be very promising. To be more specific, the ratio of energies (x12) ("which quantifies the autonomic balance between neural control of different origin" [6]) as well as the number of small accelerations (x15) were selected in most cases. Another feature, which was involved in most of the constructed features (when SMOTE was used) was the LTI (x6) which is a means to evaluate long term variability [7]. This feature alone has also relative high "usefulness" measured both in terms of FDR and AUC. Surprisingly the features quantifying the presence of decelerations were not selected so often (at least as often as we expected during the setup of the experiment) with the exception of mild decelerations which were encountered quite a few time when no synthetic samples where involved and also at 300% oversampling, and with the total duration of decelerations occurring several times for 400% oversampling. A final remark considers the appearance of the std (x2) feature which, although was quite promising at first sight, it only appeared a few times in the imbalanced case and when 500% SMOTE took place. However, the sum of

occurrences of this feature and the feature x6 is relatively constant and this can be a sign of correlation of those features (as a matter of fact they both quantify heart rate variability).

Investigating the created features as a set and not individually using the following criterion based on scatter matrices [40]:

$$J = \text{trace}\{S_w^{-1}(S_w + S_b)\} \qquad (3)$$

where $S_w = \sum_{i=1}^{2} P_i S_i$ is the within-class scatter matrix, with $S_i = E[(x - \mu_i)(x - \mu_i)^T]$ the covariance matrix of class ω_i and P_i its corresponding apriori probability and $S_b = \sum_{i=1}^{2} P_i(\mu_i - \mu_0)(\mu_i - \mu_0)^T$ is the between-class scatter matrix, where $\mu_0 = \sum_{i=1}^{2} P_i \mu_i$ is the global mean vector [40], for most of the cases there is an increase of the J value.

This change is more evident when we implement the SMOTE stage, with 75% of the newly constructed feature sets having an increased J value compared to the original feature set. Again, as in the case of the individual feature measurements, since the algorithm was not designed to maximize this criterion and since the MLP adds another stage of feature mapping through its hidden layer, not every newly constructed feature set had better separability in terms of this measure.

Finally we must also note that the grammatical evolution approach did not create too complicated expressions. This is mainly due to the fact that we used fixed sized chromosomes and

Table 3
Features created for the case of 500% overasampling and of six features

No.	F1	F2	F3	F4	F5	F6
1	x15	(−(x6))	x4*x2/(4.90)	sin(x2)	cos(−((−(x12))))	(−x4)*(−(cos(−(x5)))) + (−((exp(−(x4)))))
2	x9	exp(x15)	exp(x2−9.3 + (−sin(x3)))	x13	714.6	(−((−((−(((−(x1)))/cos(cos(x7)))))))))
3	x12	7.71	−(((sin(x13))−sin(−x2)))	x2	4.875488	log(exp(x15−19.3))/x4
4	x12	(−(x6))	cos(exp(x14/6.8))	5.5	−(x15−cos(x15)) + (x19))	exp((sin(((5.65*x4)))))
5	x19	x12 + x15	sin(−(exp(x3)))/(x9)	x6 + cos(−(−(x9)))	−exp(−((sin(x2))))	−exp(cos(x12)) + (−cos(exp(cos(x7)))))
6	x12	x2	cos(x4)	sin(sin(8.31*x8))	x15	sin(x12)/cos(((x16) + (−((−(39.1800))))))
7	exp(x15)	exp((x2−9.8))/x5	−(−(−(−(x13))))*sin(x6)	((−x14)/cos(x6))	log(exp(x3−(−log(x2))))	−(−(exp(x14)))
8	x9	exp(x15)	(exp((x2−9.28)))	(−(−(−(439.3))))	(−(−(exp(x15))))	−×18/(cos(x7)/x1)
9	x12	x6	(−((exp((x18)))))	x15 + 0.579	cos(−(x9))/sin(x8)	(−cos(x12))*exp(sin(x4/(−(cos(x12)))))
10	x2	x12	sin((x1) + 53.2) + x5	4.88	sin(sin(379.0*x12))	(−(7.7845))−log(x4)*(−(7.7845))

Table 4
Best results achieved with the proposed method, using four to six constructed features and three different values of oversampling

#Features	Overall accuracy (%)	Accuracy − (%)	Accuracy + (%)	Geometric mean (%)	SMOTE (%)
4	88.13 (0.71)	97.69 (0.50)	46.67 (2.22)	63.60 (2.40)	0
5	88.13 (0.44)	98.46 (0.31)	43.33 (2.15)	60.98 (2.33)	
6	89.38 (0.49)	98.46 (0.30)	50.00 (2.70)	65.11 (2.58)	
4	76.88 (1.13)	80.00 (0.99)	63.33 (3.16)	66.78 (2.65)	
5	85.00 (0.85)	90.00 (0.85)	63.33 (2.78)	71.12 (2.53)	300
6	81.25 (0.69)	81.54 (0.79)	80.00 (2.22)	79.68 (1.18)	
4	83.13 (0.69)	83.85 (0.73)	80.00 (1.64)	81.44 (0.93)	
5	80.00 (0.92)	80.77 (0.99)	76.66 (3.01)	74.11 (2.58)	400
6	78.13 (1.06)	79.23 (1.25)	73.33 (2.01)	75.14 (1.23)	
4	89.38 (0.69)	90.00 (0.78)	86.67 (1.64)	87.81 (0.93)	
5	90.00 (0.64)	90.00 (0.60)	90.00 (1.54)	89.68 (0.93)	500
6	88.13 (1.03)	86.92 (1.15)	93.33 (1.34)	89.73 (1.01)	

as mentioned before because the use of the MLP adds another non-linearity "element" through the neurons' activation function. In Table 3, we summarize the features that were created for the case of 500% oversampling for the case of six features.

6. Experimental results

In order to test the efficiency of our method, we compared the results of the proposed approach with the results derived from three well-known conventional methods; the k-nearest neighbor, the linear and the quadratic classifier [52] applied to the same data set. To have better results for the conventional methods and a more fair comparison, we combined these conventional methods with a dimensionality reduction stage based on principal component analysis (PCA). PCA is a very common method for the reduction of the dimensionality of the feature vector. PCA projects the original vectors onto a lower dimensional space whose axes are defined by the eigenvectors
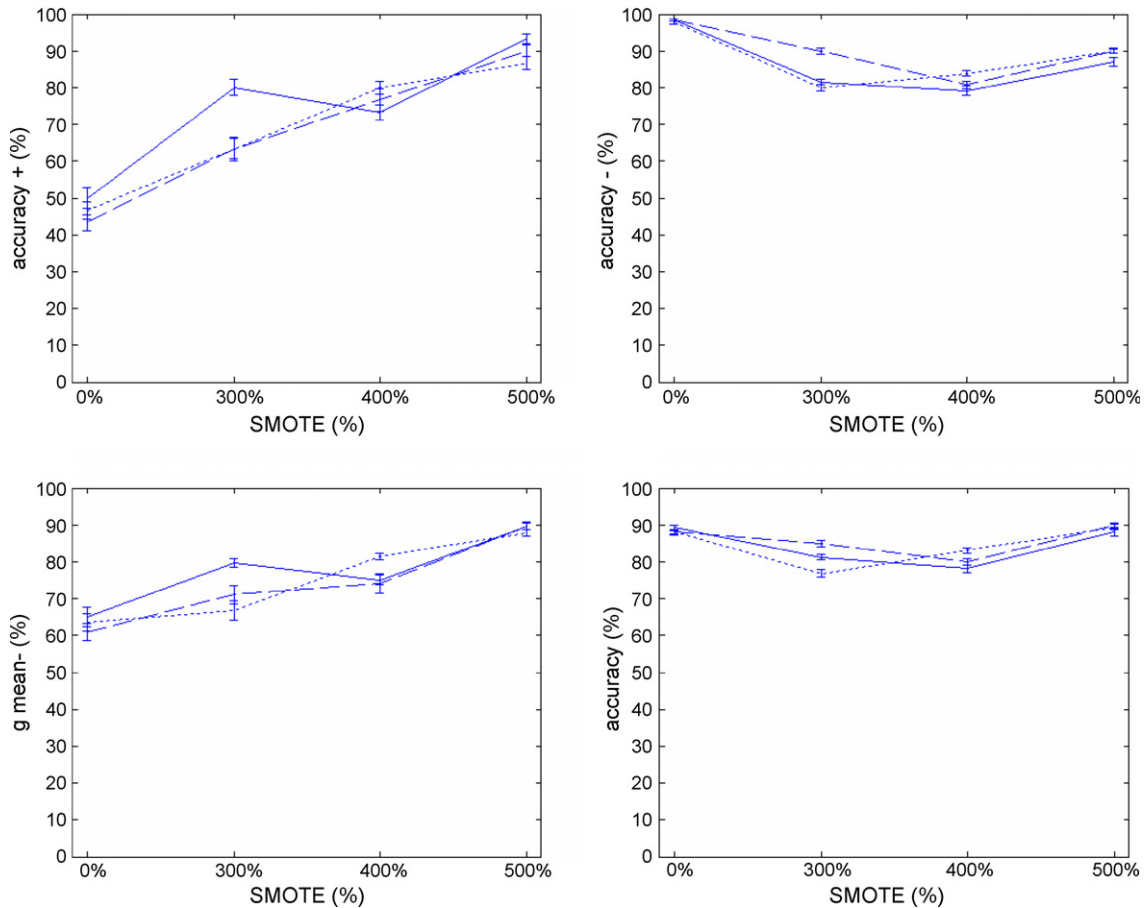

Fig. 8. The four different calculated metrics are depicted along with the 95% confidence intervals. The dotted line corresponds to four features, the dashed line to five features and the solid line to six features. The influence of the SMOTE stage is more evident when looking at the g mean graph while the accuracy + graph illustrates the positive effect that has in sensitivity the use of more constructed features.

Table 5
Best results achieved using the three conventional classifiers (with a PCA reduction stage) and the proposed method

	Overall accuracy (%)	Accuracy − (%)	Accuracy + (%)	Geometric mean (%)
Ldc (9PC)	68.75 (1.51)	66.92 (1.39)	76.67 (2.15)	71.52 (1.73)
Qdc (4PC)	72.50 (1.17)	73.85 (1.30)	66.67 (2.12)	68.98 (1.38)
1-nn (17PC) (%)	80.63 (1.14)	83.85 (1.01)	66.67 (3.35)	70.10 (2.76)
Proposed method	88.13 (1.03)	86.92 (1.15)	93.33 (1.34)	89.73 (1.01)

that correspond to the largest eigenvalues of the covariance matrix of the data [32,33].

Due to the imbalanced nature of the data set, accuracy (overall classification rate) is not the best metric choice. For example, a classifier that classifies everything negative (all cases classified as normal) will be 81.25% accurate but it will be completely useless. A more appropriate metric is the geometric mean $g = \sqrt{a^+ a^-}$ [53] (g mean), where a^+ is the accuracy, which is observed separately on positive examples (also known with the term sensitivity) and a^- is the accuracy, observed separately on negative examples (also known with the term specificity).

We experimented with four to six constructed features and three different values of oversampling and the results are summarized in Table 4 (mean values along the calculated standard error in the parentheses [54]). For comparison reasons, we also included the results of the aforementioned procedure when the original data, without the application of the SMOTE stage, were employed.

As it can be seen from Table 4, for all oversampling setups but one, the mean value of the performance metric (g mean) increases when more constructed features are added. Moreover, as we increase the percentage of the synthetically created features, the sensitivity increases reaching its maximum value for 6 constructed features and 500% oversampling. On the other hand, specificity fluctuates around 90% with its best value achieved in the case of no oversampling, which results in a heavily biased classifier. Fig. 8 depicts the results for the four different metrics (accuracy sensitivity, specificity and g mean)
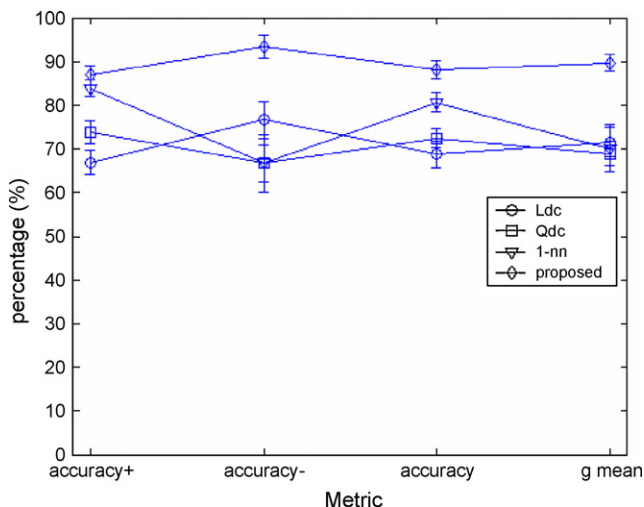

Fig. 9. The performance of the conventional classifiers compared to the proposed method.

along with the 95% confidence intervals [54]. The positive effect of the integration of the SMOTE technique in the classification process can clearly be seen.

The best results of the proposed method are achieved for the case of six selected features (even though the results with the five features could be used since there is not statistical significant difference between the two cases at a level of significance $p = 0.05$). All these results along with the corresponding results for the three conventional classifiers (for the optimal number of principal components) are summarized in Table 5. The conventional classifiers achieved worst results compared to the proposed method especially in terms of sensitivity and g mean. Fig. 9 depicts the results for the four methods in terms of accuracy, sensitivity, specificity and g mean. As it can be seen, the proposed method clearly outperforms the conventional methods used with the PCA preprocessing stage.

7. Conclusions and future work

In this work, we introduced a novel hybrid method to construct new artificial features for the FHR classification problem. The new artificial features were constructed using grammatical evolution and were tested using a neural network, which was trained based on a hybrid method. The grammatical evolution based method constructs features that give ''optimal'' results for the problem at hand, without trying to maximize the variance or the information of the features like other methods (which does not necessarily lead to better classification results). To overcome the inherent problem concerning the imbalanced distribution of the data set we adopted the SMOTE procedure which turned out to be a very effective technique. Actually, the results achieved by the conventional classifiers outperform the results achieved in other works [35] where a SMOTE stage was not involved.

The proposed method overwhelms the approach with the conventional classifiers and the PCA stage. For the ''at risk'' group the observed accuracy (a^+) was 90% for five constructed features and 93% for six constructed features with a small degradation in specificity. Especially for the case of five constructed features and 500% oversampling, the proposed procedure resulted in a totally balanced and quite high performance.

Clinical comparison of our results with those from other authors should be performed with caution, as the design of our experimental work, using two distinct pH thresholds to separate the two typical classes of normal and acidemic fetuses, was not used in other similar works concerning the prediction of

metabolic acidosis during the second stage of labor [23,24]. With the mentioned comparison limitations in mind, we can say that the results are similar to those reported in [23,24], with Chung et al. [23] achieving sensitivity equal to 87.5% and specificity equal to 75% using a pH cutoff value of 7.15, and Salamalekis et al. [24] obtaining sensitivities equal to 83.3, 100 and 100% and specificities of 97.9, 92.6 and 86.2% for pH cutoff values of 7.2, 7.15 and 7.1, respectively. However, it must be mentioned that in the work of Salamalekis et al. [24] features were also extracted from the oximetry signal, which may account for the high values in the performance of their proposed methodology.

The results of the proposed method are better compared to our previous work, where a Support Vector Machine (SVM) classifier and a PCA reduction stage was used for the same FHR features extracted from the frequency and time domain and for the same, but one, morphological features (number of small accelerations) [35]. The current results, which were achieved using an even more heavily imbalanced data set compared to other previous works [26,35], manifest the potential of the proposed method. However, further investigation is required in order to claim that this methodology is the best choice for the identification of metabolic acidosis based solely on FHR traces. In addition to this, the whole procedure has to be tested for greater number of data in order to allow "extrapolation" for any unseen case.

In this work, the classification accuracy was used as the fitness function for the selection of the features. In future work, we will examine a fitness function that is not depended on the selection of the classification method and we will try to maximize the separability of the classes in feature space [55,56]. This approach can be combined with simpler classifiers reducing the potential overfitting problems that can occur when neural networks are involved. An interesting alternative would also be to employ as the induction method SVMs which are well known for their intrinsic ability to exhibit increased generalized performance.

Other areas of interest will include increasing the number of FHR features and employing features extracted using the discrete wavelet transform [26] and probably some other morphological FHR features, namely those extracted by commercially available systems, in current use in clinical practice.

Especially, the current study showed a "reluctance" of the grammatical evolution method to select features describing decelerations. However, in clinical practice decelerations are considered to convey valuable prognostic information, so maybe a refinement of the morphological feature extraction protocol is needed. This could be done by the use of a system such as SisPorto® (Speculum, Portugal), which closely follows the FIGO guidelines for fetal monitoring [19], integrated in the central monitoring station Omniview® (Speculum, Portugal), which can also acquire data from the STAN® system for fetal electrocardiographic analysis (Neoventa, Gothemburg, Sweden) [57,58].

Finally, the Apgar score, another important clinical indicator of the fetal condition at delivery, which was not used in this study, but has been widely used by other authors [59], should also be used in future research, as another index component for the formation of classes.

Acknowledgements

References

[1] M.C. Carter, Present day performance qualities of cardiotocographs, Br. J. Obstet. Gynaecol. 100 (Suppl. 9) (1993) 10–14.

[2] H.P. Geijn, Developments in CTG analysis, Bailliere's Clin. Obstet. Gynaecol. 10 (2) (1996) 185–209.

[3] D. MacDonald, A. Grant, M. Sheridan-Pereira, P. Boylan, I. Chalmers, The Dublin randomized controlled trial of intrapartum fetal heart rate monitoring, Am. J. Obstet. Gynaecol. 152 (5) (1985) 524–539.

[4] J. Bernardes, A. Costa-Pereira, D. Ayres-de-Campos, H.P. Geijn, L. Pereira-Leite, Evaluation of interobserver agreement of cardiotocograms, Int. J. Gynaecol. Obstet. 57 (1) (1997) 33–37.

[5] D. Ayres-de-Campos, J. Bernardes, A. Costa-Pereira, L. Pereira-Leite, Inconsistencies in classification by expert of cardiotocograms and subsequent clinical decision, Br. J. Obstet. Gynaecol. 106 (12) (1999) 1307–1310.

[6] M.G. Signorini, G. Magenes, S. Cerutti, D. Arduini, Linear and non-linear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings, IEEE Trans. Biomed. Eng. 50 (3) (2003) 365–374.

[7] D. Arduini, G. Rizzo, G. Piana, A. Bonalumi, P. Brambilla, C. Romanini, Computerized analysis of fetal heart rate: I. Description of the system (2CTG), J. Matern. Fetal Invest. 3 (1993) 159–163.

[8] R. Mantel, H.P. Geijn, F.J. Caron, J.M. Swartjes, E.E. Woerden, H.W. Jongsma, Computer analysis of antepartum fetal heart rate: 1. Baseline determination, Int. J. Biomed. Comput. 25 (4) (1990) 261–272.

[9] R. Mantel, H.P. Geijn, F.J. Caron, J.M. Swartjes, E.E. Woerden, H.W. Jongsma, Computer analysis of antepartum fetal heart rate: 2. Detection of accelerations and decelerations, Int. J. Biomed. Comput. 25 (4) (1990) 273–286.

[10] M. Mongelli, R. Dawkins, T. Chung, D. Sahota, J.A.D. Spencer, A.M.Z. Chang, Computerized estimation of the baseline fetal heart rate in labour: the low frequency line, Br. J. Obstet. Gynaecol. 104 (10) (1997) 1128–1133.

[11] G.M. Taylor, G.J. Mires, E.W. Abel, S. Tsantis, T. Farrell, P.F. Chien, Y. Liu, The development and validation of an algorithm for real time computerized fetal heart rate monitoring in labour, Br. J. Obstet. Gynaecol. 107 (9) (2000) 1130–1137.

[12] G.S. Dawes, M. Moulden, C.W. Redman, Computerized analysis of antepartum fetal heart rate, Am. J. Obstet. Gynaecol. 173 (4) (1995) 1353–1354.

[13] J. Jezewski, J. Wrobel, Fetal monitoring with automated analysis of cardiotocogram: the KOMPOR system, in: Proceedings 15th Ann Card IEEE/EMBS, San Diego, CA, (1993), pp. 638–639.

[14] K. Maeda, Computerized analysis of cardiotocograms and fetal movements, Bailliere's Clin. Obstet. Gynaecol. 4 (4) (1990) 797–813.

[15] C. Ulbricht, G. Dorffner, A. Lee, Neural networks for recognizing patterns in cardiotocograms, Artif. Intell. Med. 12 (3) (1998) 271–284.

[16] O. Fontenla-Romero, A. Alonso-Betanzos, B. Guijarro-Berdinas, Adaptive pattern recognition in the analysis of cardiotocographic records, IEEE Trans. Neural Netw. 12 (5) (2001) 1188–1195.

[17] J. Bernardes, C. Moura, J.P.M. Sa, L. Pereira-Leite, The Porto system for automated cardiotocographic signal analysis, J. Perinat. Med. 19 (1–2) (1991) 61–65.

[18] J. Bernardes, C. Moura, J.P.M. Sa, L. Pereira-Leite, H.P. Geijn, The Porto system, in: H.P. van Geijn, F.J.A. Copray (Eds.), A Critical Appraisal of Fetal Surveillance, Elsevier Science, New York, 1994, pp. 315–324.

[19] D. Ayres-de-Campos, J. Bernardes, A. Garrido, J.P.M. Sa, L. Pereira-Leite, SisPorto 2.0: a program for automated analysis of cardiotocograms, J. Matern. Fetal Med. 9 (5) (2000) 311–318.

[20] G. Magenes, M.G. Signorini, D. Arduini, Classification of cardiotocographic records by neural networks, in: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00), vol. 3, 2000, pp. 637–641.

[21] G. Magenes, M.G. Signorini, R. Sassi, D. Arduini, Multiparametric analysis of fetal heart rate: comparison of neural and statistical methods, Proc. Medicon (2001), pp. 360–363.

[22] S. Kol, I. Thaler, N. Paz, O. Shmueli, Interpretation of nonstress tests by an artificial neural network, Am. J. Obstet. Gynaecol. 172 (5) (1995) 1372–1379.

[23] T.K.H. Chung, M.P. Mohajer, X.J. Yang, A.M.Z. Chang, D.S. Sahota, The prediction of fetal acidosis at birth by computerized analysis of intrapartum cardiotocography, Br. J. Obstet. Gynaecol. 102 (6) (1995) 454–460.

[24] 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, Br. J. Obstet. Gynaecol. 109 (10) (2002) 1137–1142.

[25] Z.R. Struzik, W.J. Wijngaarden, Cumulative effective Hölder exponent based indicator for real time fetal heartbeat analysis during labour, Report INS-R0110 2001.

[26] 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. AI Tools 15 (2006) 411–432.

[27] E.C. Ifeachor, R.D.F. Keith, J. Westgate, K.R. Greene, An expert system to assist in the management of labour, in: Proceedings World Congress on Expert Systems, vol. 4, 1991, pp. 2615–2622.

[28] J.F. Skinner, J.M. Garibaldi, E.C. Ifeachor, A fuzzy system for fetal heart rate assessment, in: Proceedings 6th Fuzzy Days on Computational Intelligence, Dortmund, Germany, (1999), pp. 20–29.

[29] A. Alonso-Betanzos, V. Moret-Bonillo, L.D. Devoe, J.R. Searle, B. Banias, E. Ramos, Computerized antenatal assessment: The NST-EXPERT project, Automedica 14 (1992) 3–22.

[30] A. Alonso-Betanzos, B. Guijarro-Berdinas, V. Moret-Bonillo, S. Lopez-Gonzalez, The NST-EXPERT project: the need to evolve, Artif. Intell. Med. 7 (4) (1995) 297–313.

[31] B. Guijarro-Berdinas, A. Alonso-Betanzos, O. Fontenla-Romero, Intelligent analysis and pattern recognition in cardiotocographic signals using a tightly coupled hybrid system, Artif. Intell. 136 (2002) 1–27.

[32] S. Haykin, Neural Networks: A Comprehensive Foundation, second ed., Prentice Hall, Englewood Cliffs, NJ, 1999.

[33] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, New York, 1995.

[34] G. Georgoulas, C. Stylios, J. Bernardes, P.P. Groumpos, Classification of cardiotocograms using Support Vector Machines, in: Proceedings 10th IFAC Symposium on Large Scale Systems: Theory and Applications (LSS'04), July 26–28, 2004, Osaka, Japan, 2004.

[35] G. Georgoulas, C. 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 (5) (2006) 875–884.

[36] M. O'Neill, C. Ryan, Grammatical Evolution, IEEE Trans. Evol. Comput. 5 (4) (2001) 349–358.

[37] J.R. Koza, On the Programming of Computers by Means of Natural Selection, MIT Press, Cambridge, MA, 1992.

[38] P. Naur, Revised Report on the Algorithmic Language ALGOL 60, Commun. ACM 6 (1) (1963) 1–17.

[39] I.G. Tsoulos, D. Gavrilis, E. Dermatas, Features Selection and Construction using Grammatical Evolution, submitted to IEEE Transactions on Evolutionary Computation.

[40] S. Theodoridis, K. Koutroumbas, Pattern Recognition, third ed., Elsevier, 2006.

[41] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1/2) (1997) 273–324.

[42] J.P.M. Sa, Pattern Recognition, Concepts, Methods and Applications, Springer, 2001.

[43] G. Rooth, A. Huch, R. Huch, Guidelines for the use of fetal monitoring, Int. J. Gynaecol. Obstet. 25 (1987) 159–167.

[44] National Institute of Child Health and Human Development Research Planning Workshop, Electronic fetal heart rate monitoring: research guidelines for interpretation, Am. J. Obstet. Gynaecol. 177 (6) (1997) 1385–1390.

[45] H. Goncalves, A.P. Rocha, D. Ayres-de-Campos, J. Bernardes, Internal versus external intrapartum foetal heart rate monitoring: the effect on linear and nonlinear parameters, Physiol. Meas. 27 (2006) 307–319.

[46] N.V. Chawla, N. Japkowicz, A. Kotcz, Editorial: special issue on learning from imbalanced data sets, SIGKDD Explor. 6 (1) (2004) 1–6.

[47] N.V. Chawla, L. Hall, W. Kegelmeyer, SMOTE: Synthetic Minority Over-Sampling Technique, J. Artif. Intell. Res. 16 (2002) 3241–3357.

[48] M.J.D. Powell, A tolerant algorithm for linearly constrained optimization calculations, Math. Program. 45 (1989) 547–566.

[49] R.L. Haupt, S.E. Haupt, Practical Genetic Algorithms, John Wiley & Sons, New Jersey, 2004.

[50] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth International Group, Belmont, 1984.

[51] S.L. Salzberg, On comparing classifiers: pitfalls to avoid and a recommended approach, Data Min. Know. Disc. 1 (3) (1997) 317–328.

[52] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., John Wiley & Sons, 2001.

[53] F. Provost, T. Fawcett, Analysis and visualization of classifier performance: comparison under imprecise class and cost distribution, in: Proceedings 3rd International Conference on Knowledge Discovery and Data Mining, 1997, pp. 43–48.

[54] A. Flexer, Statistical evaluation of neural network experiments: minimum requirements and current practice, Technical Report, The Austrian Research Institute of Artificial Intelligence, 1994.

[55] L.S. Burrell, S.M. Glynn, G.J. Vachtsevanos, B. Litt, Feature analysis of functional MRI for discrimination between normal and epileptogenic brain, in: Proceedings European Control Conference, Kos, Greece, (2007), pp. 2554–2559.

[56] O. Smart, H. Firpi, G. Vachtsevanos, Genetic programming of conventional features to detect seizure precursors, Engineering Applications of Artificial Intelligence, in press.

[57] I. Amer-Wahlin, et al., Cardiotocography only versus cardiotocography plus ST analysis of fetal electrocardiogram for intrapartum fetal monitoring: a Swedish randomised controlled trial, Lancet 358 (9281) (2001) 534–538.

[58] K.G. Rosén, Intrapartum fetal monitoring and the fetal ECG—time for a change, Arch. Perinat. Med. 7 (2001) 7–12.

[59] J. Bernardes, D. Ayres-de-Campos, A. Costa-Pereira, L. Pereira-Leite, A. Garrido, Objective computerized fetal heart rate analysis, Int. J. Gynaecol. Obstet. 62 (2) (1998) 141–147.