Fetal Heart Rate Classification: First vs. Second Stage of Labor

J Spilka[1], R Leonarduzzi[2], V Chudáček[1], P Abry[2], M Doret[3]

[1]CIIRC, Czech Technical University in Prague, Czech Republic;
[2]Univ Lyon, Ens de Lyon, Univ Claude Bernard, CNRS, Laboratoire de Physique, F-69342 Lyon, France;
[3]Femme-Mère-Enfant Hospital, Bron, France;

Abstract

Fetal Heart Rate (FHR) is clinically used for early detection of fetal acidosis. Despite a marked interest in automatic detection procedures, FHR analysis remains a challenging signal processing task, receiving considerable research attention. Among other difficulties, the two stages of labor (dilation and active pushing) produce very different FHR dynamics. Most research efforts, however, have either ignored these differences or analyzed only one of the two stages of labor. In this work, we propose to assess the impact of labor stages on acidosis detection performance. A state-of-the-art sparse support vector machine classifier that performs simultaneously feature selection and classification is applied to a large-size and well documented FHR database. It shows that the selected set of features differs for each stage and that detection performance improves when the difference between labor stages is considered.

Keywords *Fetal Heart Rate, Acidosis detection labor stages, scale-free features, Sparse SVM,*

1 Introduction

Intrapartum Fetal Monitoring. Fetal heart rate (FHR) provides major information about fetal health and is thus routinely monitored in clinical practice. It is mainly used to assess well-being of the fetus, and to decide on an operative delivery. In daily clinical routine, FHR is examined by visual inspection following clinical guidelines issued by the International Federation of Gynecology and Obstetrics (FIGO) [1]. However, the intrinsic complexity of FHR makes its visual interpretation difficult and the sole use of FIGO guidelines leads to a substantial inter and intra observer variability [2], which is in part responsible for a growing number of unnecessary Caesarean sections [3]. There are hence numerous research efforts devoted to automated fetal acidosis early detection.

Automatic FHR processing. Automatic acidosis detection relies on the use of supervised machine learning, based on features aiming to capture the relevant characteristics of FHR temporal dynamics. A wide range of signal processing techniques have been explored to devise such features, ranging from computerized FIGO guidelines [1] to multifractal analysis [4]; cf. [5] for review.

Labor stages. Automatic FHR analysis is further complicated by the existence of two distinct labor stages. The first stage (dilatation), is characterized by progressive cervical dilatation and regular contractions. The second stage (active pushing), is characterized by a fully dilated cervix and expulsive contractions. Both stages are characterized by largely different temporal dynamics.

State-of-the-art approach is to study either single labor stage alone, cf. e.g. [6,7] or not to distinguish between the stages at all [8,9]. While the former approach is methodologically correct, it discards data that might be useful for detection improvement. The latter approach is impaired by the potential different FHR dynamics: relevant features might thus change drastically from one stage to the other, and negatively impact classifier generalization ability.

Related works. There have been only few attempts to study the impact of the transition between stages in FHR detection. Nevertheless, some preliminary analyses have been performed to assess how each stage impacts the Hurst exponent [10] and entropy rates [11]. However there is no systematic reports on how such stage differences impact supervised classification.

Goals, contributions and outline. The present contribution aims to investigate the impact of labor stages in supervised classification. Both the selection of relevant features and classification performance are compared between the two stages, with emphasis on the existence of features that are discriminative in both stages. To achieve these goals, Sparse-Support Vector Machine (S-SVM), for joint classification and feature selection, is applied to a comprehensive set of FHR features, computed from a large FHR database (cf. Section 2). Classification performance and feature selection are compared, jointly for both stages and independently for each one, cf. Section 3.

2 Methods

Database. FHR data were collected at Femme-Mère-Enfant hospital, in Lyon, France, during daily routine monitoring from 2000 to 2010. Recording were performed using STAN S21 or S31 devices with internal scalp electrodes. Clinical information was provided by the obstetrician in charge, notably the umbilical artery

Work supported by Czech Science Foundation Agency project No. 14-28462, ANR AMATIS grant, and Hospices Civil de Lyon, Hôpital Femme Mere Enfant, Project Hospitalier de Recherche Clinique.

Table 1: **Clinical data for both stages (acidotic *vs.* normal group)**, reported as mean (standard deviation). Statistical differences between acidotic and normal subjects($p < 0.05$) are indicated with †.

	S_I		S_{II}	
	Acidotic n=27	Normal n=1015	Acidotic n=28	Normal n=734
Birth-weight (g)	3383 (446)	3329 (472)	3452 (400)	3366 (444)
Operative delivery for fetal distress (n)	12 (44%)	213 (21%)†	13 (46%)	151 (21%)†
Umbilical cord arterial pH	7.01 (0.03)	7.24 (0.07)†	7.01 (0.04)	7.22 (0.06)†
Apgar score at 5 minutes	9.37 (0.93)	9.89 (0.53)	9.57 (0.79)	9.90 (0.43)
Length of second stage (min)	8.67 (5.02)	6.82 (5.09)	27.86 (9.67)	27.64 (9.85)
Time from end of rec. and birth (min)	1.61 (2.72)	0.81 (1.87)†	1.93 (3.36)	0.96 (2.08)

pH after delivery and the decision for intervention due to suspected acidosis [12]. Subject inclusion criteria were detailed in [7, 12], leading to a set, S, of $N = 1804$ recordings, gathering: *acidotic* subjects, $N_+ = 55$, with pH ≤ 7.05 and *normal* cases, $N_- = 1749$, with pH > 7.05 [13]. For the purpose of first *vs.* second stage comparison, subjects were further split into two groups based on the second stage duration (t_{II}): set S_I with $t_{II} \leq 15$ min and set S_{II} with $t_{II} > 15$ min. Relevant clinical data are reported in Table 1. FHR analysis was systematically conducted on the last 20 minutes of FHR recordings, as illustrated in Fig. 1.

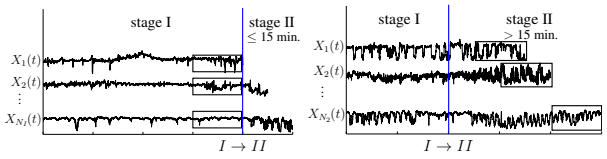

Figure 1: **Analyzed regions.** Rectangles indicate the last 20 min under analysis S_I (left) and S_{II} (right).

Features. The set of 20 features as described in [7] is used. They are organized in three groups labeled *automated FIGO*, *spectral* and *scale-free dynamics*.

FIGO enhanced and automated FIGO features. They are based on FIGO guidelines, used by obstetricians to evaluate FHR: baseline evolution, variability and characterisation of accelerations/decelerations [1]. Baseline evolution is modeled by a linear regression: $B(t) = \beta_0 + \beta_1 t$. Long and short term variability (LTV and STV, respectively) are computed with the standard procedures detailed in [1]. The number of accelerations and decelerations ($\#acc$ and $\#dec$) are counted using the definitions in [1]. Finally, decelerations are further quantified by their average depth MAD_{dtrd}, average duration T_{stress} and average area A_{dec}.

Spectral features. Spectral behavior of FHR is quantified by computing the energy in predefined frequency bands. Since no consensus has been reached on the definition of such bands for fetuses (cf. [14, 15] for discussions), the definitions for adults [15] are used: very low frequency E_{VLF} ($[0.003, 0.04]$ Hz), low frequency E_{LF} ($[0.04, 0.15]$ Hz), and high frequency E_{HF} ($[0.04, 0.15]$ Hz). Finally, the ratio of E_{LF} and E_{HF}, denoted as LF/HF, and the spectral index α [15], estimated over both LF and HF bands, are computed. All spectral estimates are computed using the Welch periodogram.

Scale-free dynamics features. Following [4, 7, 15], scale-invariance/multifractal features are computed to quantify the multiscale and complex FHR temporal dynamics. All these features are estimated using linear regressions based on relevant multiresolution quantities. Features H and h_{min} are computed from the moments of wavelet coefficients. Features c_1, c_2, c_3 and c_4 are computed from the cumulants of wavelet leaders [16]. Features H and c_1 are related to the correlation structure of FHR, while h_{min}, c_2, c_3 and c_4 measure information contained in its higher-order statistics. see also e.g. [4, 15].

Feature preprocessing. Outliers were removed by Winsorization in the interval $[Q_1 - 3IQR, Q_3 + 3IQR]$, where Q_i is the i-th quartile and $IQR = Q_3 - Q_1$ is the interquartile range. All features were standardized.

Sparse Support Vector Machine. S-SVM is a machine-learning tool that performs jointly classification and feature selection [17]. Like traditional SVM, S-SVM computes an optimal hyperplane that separates normal and acidotic cases. In addition, S-SVM performs feature selection by imposing an ℓ_1-norm regularization that leads to a decision rule that effectively involves only a limited subsets of features regarded as relevant. S-SVM thus outputs a feature-weight vector $\mathbf{w} = (w_i)$ that quantifies the importance granted to each feature: $w_i = 0$ indicates features that are poorly discriminant and thus not used in classification, whereas larger w_i indicates a large discriminative power of feature i. Training of S-SVM depends on a regularization parameter C that controls the trade-off between decision rule (or feature) sparsity and misclassification rate (with higher values of C decreasing sparsity). For further details on S-SVM, see [7, 17] and references therein.

Performance assessment. Performance is quantified by the specificity (SP), sensitivity (SE) and balanced error rate: $BER = (SP + SE)/2$. Selection of C, computation of weights w_i and performance assessment are performed using double-loop stratified k-fold cross validation (CV), where k is chosen as the number of acidotic cases (see [7] for details).

3 Comparisons between the labor stages

Pairwise correlation. Fig. 3 displays the pairwise correlations of all features, for each stage, and reveals several interesting characteristics. First, the correlation structure

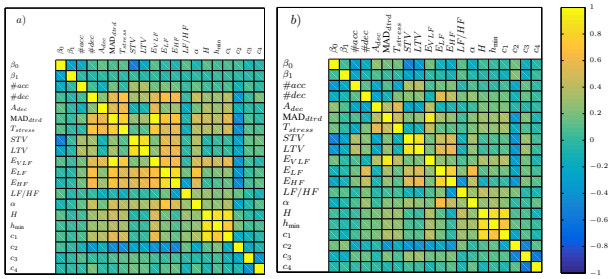

Figure 3: **Correlation.** Pairwise correlation matrix of all the features for the two labor stages: *a)* S_I, *b)* S_{II}.

Table 2: **Selected features and weights.**

S	w	S_I	w	S_{II}	w
c_1	.68	MAD_{dtrd}	.82	c_1	.89
T_{stress}	.43	β_0	.50	c_2	.45
MAD_{dtrd}	.41	c_1	.24		
c_2	.29	T_{stress}	.16		
E_{HF}	.18				
STV	.17				
β_0	.16				

is similar for both stages, but overall correlation is higher during S_I than during S_{II}. Second, several features (e.g. $\beta_0, \beta_1,, c_2$) are uncorrelated to all others, irrespective of the stage. Third, features H, c_1, and $\text{h}_{\min}$ have very high correlations [7] in both stages. Due to proximity of H, $\text{h}_{\min}$ and c_1 in nature, H and $\text{h}_{\min}$ are considered redundant to c_1 and are removed from further analysis to ease interpretation of results. Other highly-correlated features are kept since they are of different natures.

Feature selection and classification. Fig. 2 displays the weights produced for each feature (top panels) and classification performance (bottom panels), as functions of the sparsity parameter C. First, it illustrates that low values of C promote sparsity with less features involved in classification. Second, it shows that optimal performance is obtained for a level of sparsity referred to as C_{opt} that never corresponds to the use of all available features. This highlights the need to perform feature selection to prevent unnecessary over-complicated and over-fitted models.

Optimal feature set. Table 2 shows selected features and their corresponding weights, at the optimal level $C = C_{opt}$ (only those with nonzero weights are displayed). It can be seen that classification in S_I requires only four features: MAD_{dtrd} and T_{stress} (decelerations' amplitude and frequency), β_0 (baseline level), and c_1 (scale-free *linear* variability). For S_{II} the feature vector is even more sparse and contains only two features: c_1 and c_2 (scale-free *nonlinear* variability). In contrast, for S, not only the features that are significant for S_I and S_{II}, but also extra features such as measures of short term variability like STV and E_{HF} are included. These additional features are likely needed to account for the additional inter-stage variability, which leads to an overall decrease in detection performance compared to what can be achieved using the knowledge about the stage of the delivery. Finally, Table 2 highlights that c_1 is the only feature used in all groups. Interestingly, it consists of robust quantification of FHR variability (cf. e.g., [7, 15]).

First *versus* second stage. S-SVM selects for S_I features classically rooted in clinical practice, such as MAD_{dtrd}, T_{stress} (both quantifying the impact of decelerations), and β_0 (average level of baseline). Interestingly, these features are no longer used for S_{II}. Since the second stage is associated with active maternal pushing, large and frequent decelerations are present in most records, irrespective of acidosis. In light of the loss of discriminative power from MAD_{dtrd}, T_{stress}, S-SVM conveniently replaces them with c_2 (which is associated with changes in local regularity and burstiness of data), as a companion to the already selected c_1.

Optimal classification performance. Classification performance for $C = C_{opt}$ is presented in Table 3. It indicates that that independent evaluation of S_I and S_{II}

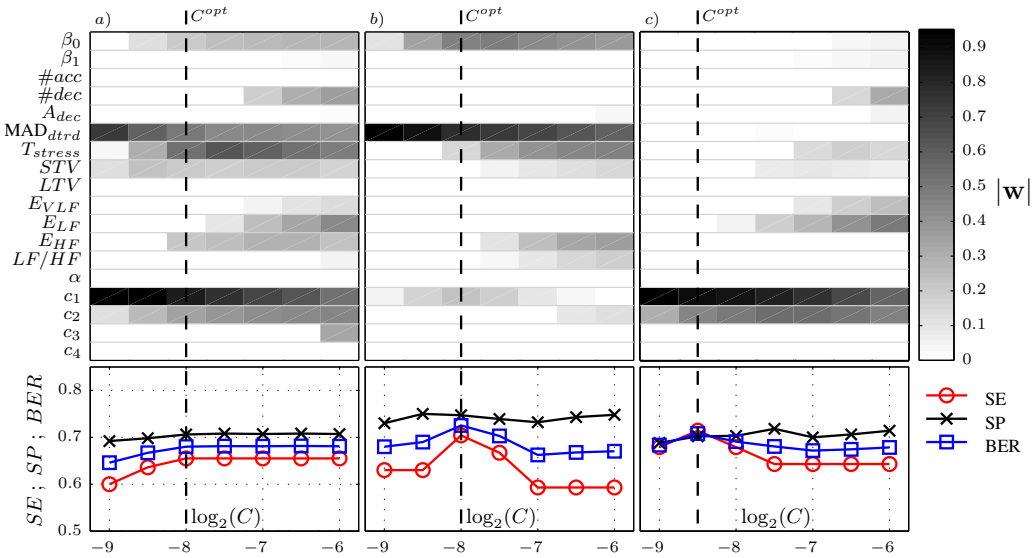

Figure 2: **S-SVM performance.** Feature selection (top row) and classification performance (bottom row) as function of the regularization parameter C. Results for different sets: *a)* S, *b)* S_I, and *c)* S_{II}.

Table 3: **Optimal classification performance** for different combinations of training / testing sets.

Tr / Te	SE	SP	BER	#TP	#FN	#TN	#FP
S/S	.62	.71	.66	34	21	1241	508
S_I/S_I	.67	.74	.70	18	9	752	263
S_{II}/S_{II}	.68	.70	.69	19	9	515	219
S/S_I	.56	.80	.68	15	12	807	208
S/S_{II}	.68	.59	.64	19	9	434	300

results in better classification performance, since simple models with only the relevant features for each stage are used. Interestingly, if the classifier trained from S is tested only with samples from S_I and S_{II}, dramatic loses of either sensitivity or specificity are observed; this indicates that the loss of performance is due to a suboptimal training of the classifier that fails to fully account for the characteristics of each stage.

4 Conclusions

This contribution explores the influence of the two stages of labor on feature selection and classification performance in a supervised classification task. To that end, it uses a comprehensive set of FHR features and a Sparse-SVM framework on three scenarios: i) naive classification without recognizing labor stages ; ii) separate classification of records in the first stage ; iii) separate classification of the records in the second stage. It was shown that failure to recognize the stages leads to a complex model, involving a large number of features, with inferior performance. In contrast, results indicate that an independent evaluation of both stages provides simpler models (less features) with better performance. Further, selected features for the first stage confirm the predominance of decelerations and variability for acidosis detection while, for the second stage, decelerations are no longer informative and other measures of variability, namely c_1 and c_2, are preferred.

References

[1] D. Ayres-de Campos, et al. Figo consensus guidelines on intrapartum fetal monitoring: Cardiotocography. *Int J Gynaecol Obstet*, 131(1):13–24, Oct 2015.

[2] L. Hruban, et al. Agreement on intrapartum cardiotocogram recordings between expert obstetricians. *Journal of Evaluation in Clinical Practice*, 21(4):694–702, 2015.

[3] Z. Alfirevic, D. Devane, and G. Gyte. Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. *Cochrane Database Syst Rev*, 3(3):CD006066, 2006.

[4] M. Doret, et al. Multifractal analysis of fetal heart rate variability in fetuses with and without severe acidosis during labor. *Am J Perinatol*, 28(4):259, 2011.

[5] J. Spilka, et al. Using nonlinear features for fetal heart rate classification. *Biomed Signal Process Control*, 7(4):350–357, 2012.

[6] A. Georgieva, et al. Phase-rectified signal averaging for intrapartum electronic fetal heart rate monitoring is related to acidaemia at birth. *BJOG*, 121(7):889–894, Jun 2014.

[7] J. Spilka, et al. Sparse support vector machine for intrapartum fetal heart rate classification. *IEEE J Biomed and Health Inform*, PP(99):1–1, 2016.

[8] A. Costa, et al. Prediction of neonatal acidemia by computer analysis of fetal heart rate and st event signals. *Am J Obstet Gynecol*, 201(5):464.e1–464.e6, Nov 2009.

[9] P.A. Warrick, et al. Classification of normal and hypoxic fetuses from systems modeling of intrapartum cardiotocography. *IEEE Trans Biomed Eng*, 57(4):771–779, 2010.

[10] J. Spilka, et al. Impacts of first and second labour stages on hurst parameter based intrapartum fetal heart rate analysis. In *Computing in Cardiology Conference (CinC), 2014*, pages 777–780, 2014.

[11] Jongil Lim, et al. Quantitative comparison of entropy analysis of fetal heart rate variability related to the different stages of labor. *Early Hum Dev*, 90(2):81–85, Feb 2014.

[12] M. Doret, et al. Use of peripartum st analysis of fetal electrocardiogram without blood sampling: a large prospective cohort study. *Eur J Obstet Gynecol Reprod Biol*, 156(1):35–40, May 2011.

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

[14] S. Siira, et al. Do spectral bands of fetal heart rate variability associate with concomitant fetal scalp pH? *Early Hum Dev*, 89(9):739–742, Sep 2013.

[15] M. Doret, et al. Fractal Analysis and Hurst Parameter for intrapartum fetal heart rate variability analysis: A versatile alternative to Frequency bands and LF/HF ratio. *PLoS ONE*, 10(8):e0136661, 08 2015.

[16] H. Wendt, P. Abry, and S. Jaffard. Bootstrap for Empirical Multifractal Analysis. *IEEE Signal Proc. Mag.*, 24(4):38–48, 2007.

[17] L. Laporte, et al. Nonconvex Regularizations for Feature Selection in Ranking With Sparse SVM. *IEEE Trans Neural Netw Learn Syst*, 25(6):1118–1130, 2014.

Address for correspondence:

Jiří Spilka, jiri.spilka@ciirc.cvut.cz

CIIRC, Czech Technical University in Prague