Sparse Support Vector Machine for Intrapartum Fetal Heart Rate Classification

Jiří Spilka[1,2], Jordan Frecon[1], Roberto Leonarduzzi[1], Nelly Pustelnik[1], Patrice Abry[1], Muriel Doret[3]

***Abstract*—Fetal Heart Rate (FHR) monitoring is routinely used in clinical practice to help obstetricians assess fetal health status during delivery. However, early detection of fetal acidosis that allows relevant decisions for operative delivery remains a challenging task, receiving considerable attention. This contribution promotes Sparse Support Vector Machine (SVM) classification that permits to select a small number of relevant features and to achieve efficient fetal acidosis detection. A comprehensive set of features is used for FHR description, including enhanced and computerized clinical features, frequency domain, and scaling and multifractal features, all computed on a large (1288 subjects) and well documented database. The individual performance obtained for each feature independently is discussed first. Then, it is shown that the automatic selection of a sparse subset of features achieves satisfactory classification performance (sensitivity 0.73 and specificity 0.75, outperforming clinical practice). The subset of selected features (average depth of decelerations $\mathbf{MAD}_{dtrd}$, baseline level β_0, and variability H) receive simple interpretation in clinical practice. Intrapartum fetal acidosis detection is improved in several respects: A comprehensive set of features combining clinical, spectral and scale-free dynamics is used; an original multivariate classification targeting both sparse feature selection and high performance is devised; state-of-the-art performance is obtained on a much larger database than that generally studied with description of common pitfalls in supervised classification performance assessments.**

***Index Terms*—Biomedical Signal Processing, Supervised Classification, Sparse SVM, Feature Selection, Fetal Heart Rate**

I. Introduction

Intrapartum fetal monitoring. Fetal heart rate (FHR) is routinely monitored to assist obstetricians with the evaluation of the oxygenation status of fetuses during delivery. Notably, they are used for early fetal acidosis detection and hence for taking timely and relevant decisions for operative deliveries, to prevent adverse asphyxia outcomes, such as neural development disability, neonatal encephalopathy, and cerebral palsy [1]. In daily clinical practice, FHR is mostly examined visually following guidelines edited by national and international scientific societies, such as the International Federation of Gynecology and Obstetrics (FIGO) [2], that mostly focus on decelerations and accelerations (frequency of occurrence, shape and depth, synchronisation with uterine contractions), long term variabilities, and baseline levels and trends. However, it is well known that FHR is regulated by multiple and diverse neurological feedback loops (baroreceptors, chemoreceptors), by hormones, and influenced with various external factors, such as infections, resulting into complex temporal dynamics, that can not be easily accounted for by criteria used under visual inspection. Such complexities result in large inter and even intra-observer variabilities in FHR analysis, cf. e.g. [3], and are also believed to be the direct cause for a high rate of cesarean sections for suspected fetal acidosis, that are a posteriori found to have been unnecessary [4]. These issues have triggered numerous research efforts to enhance and automatize FHR analysis and fetal acidosis early detection.

Related works. While FIGO criteria mostly focuses on morphological properties of FHR [5], a set of features, referred to as linear, has been designed to enhance automatic FHR analysis by quantifying temporal dynamics, via concepts such as autocorrelation and spectral analysis, and models that either rely on characteristic time scales [6] or on scale-free paradigms [7], [8]. Further, features designed to quantify FHR complexities, referred to as nonlinear (because they measure temporal dynamics beyond correlation), and mostly based on variations of entropy rates [9], were also used. The recently proposed multifractal and scattering analyses also marry nonlinear complexity measures with scale-free dynamics [7], [10].

Assessment of the performance of each feature for FHR analysis and acidosis detection is retrospective and relies mainly on pH (measured from post-birth umbilical cord artery blood sample), used as a marker of fetal acidosis, thus leading to focus FHR analysis on the minutes preceding delivery. Results reported in the literature indicate that fetal acidosis can be equally well detected using various independent features [6], [11]–[13], measuring different aspects of FHR. However, they also reveal that performance obtained from single (or univariate) features is not satisfactory enough to be transfered to clinical practice, thus prompting for the joint (or multivariate) use of several features. The recourse to supervised classification and machine learning has thus been investigated, cf. [13]–[20]. Despite promising results, contributions available in the literature suffer from classical limitations: Generally, small size databases (from a few tens to a few hundreds of subjects, with the notable exception of [17], [21] using several thousands) versus large number of features,

[1] Univ Lyon, Ens de Lyon, Univ Claude Bernard, CNRS, Laboratoire de Physique, F-69342 Lyon, France
patrice.abry@ens-lyon.fr, firstname.lastname@ens-lyon.fr
[2] CIIRC, Czech Technical University in Prague, Czech Republic
jiri.spilka@ciirc.cvut.cz
[3] Femme-Mère-Enfant Hospital, Lyon, France

Work supported by ANR 2010 French grants #18535 FETUSES and AMATIS #112432, by #NT11124-6/2010, and by Hospices Civil de Lyon's PHRC. Neoventa Medical is gratefully acknowledged for assistance with STAN technologies. C. Stylios and LKIC, TEI, Greece, are gratefully acknowledged for sharing parallel computing resources.

Published version DOI: 10.1109/JBHI.2016.2546312

resulting in a lack of robustness and generalizability [18]; Complicated nonlinear feature transformations or intricate decision boundaries, impairing interpretation and thus transfer to clinical practice [16], [17].

Outline, goals, and contributions. In that context, the present contribution aims to explore the benefits of imposing *sparsity* into supervised classification for fetal acidosis detection based on the joint use of a collection of features. Its originality is twofold: First, it makes use of a large (1288 subjects) and well-documented database of FHR data, collected at a French public academic hospital (cf. Section II); Second, it constructs the classification procedure from a Sparse Support Vector Machine (S-SVM) technique that permits simultaneously to achieve optimal classification performance and to select, within a long list, a small number of features that actually contribute to classification (cf. Section III for feature description, gathering FIGO enhanced and automated, spectral and scale-free multifractal features, and S-SVM principles and algorithms). Results, discussed in Section IV, show the need to account for imbalanced class sizes (healthy versus acidotic subjects) and the importance of a careful design of performance assessment procedures (double loop cross validation) to avoid biased evaluations. Most importantly, the relevance and impact of sparsity in combining a large set of features for joint fetal acidosis detection is illustrated, quantified and discussed with respect to model complexity and ease of interpretation.

II. Material: database and preprocessing

Database. FHR data were collected at the public academic French Hospital Femme-Mère-Enfant (Lyon, France), from 2000 to 2010, during daily routine monitoring. Data were recorded using STAN S21 or S23 devices via internal fetal scalp electrodes. The database consists of 3049 recordings, each systematically documented by obstetricians in charge of delivery. It also contains clinical information, notably providing pH after delivery and obstetrician decision for intervention because of diagnosed fetal acidosis, cf. [22] for details.

A previous contribution by the authors reported significant differences between the statistics of the temporal dynamics of the first (dilatation) and second stages of labor [23]. To permit a clear assessment of the proposed classification and feature selection strategies, not blurred by mixing the differences stemming from the two stages, it has thus been chosen to perform analysis on the first stage only. Subject inclusion criteria are [22]: gestational age $\geq$ 37 (weeks) and maternal age $\geq$ 18. Further, data quality requirements are: presence of arterial pH and physiological difference to venous pH, signal quality (length $\geq$ 30 min, amount of missing data $\leq$ 15%, time between end of recording and birth $\leq$ 10 min), continuous monitoring within first stage ending less than 20 minutes before delivery (and pH measurement). The criteria led to a subset of $N = 1288$ selected recordings, which were further divided into two classes: *acidotic*, $N_+ = 37$, with pH ≤ 7.05 and *normal*, $N_- = 1251$, with pH > 7.05 [24]. Clinical data are reported in Table I. The available documentation permitted to compute the clinical benchmark performance, as reported in Table II, showing satisfactory specificity (SP, % of correctly identified negative cases) but low sensitivity (SE, % of correctly identified positive cases). The number of true positives, false positives, false negatives and true negatives (#TP, #FP, #FN and #TN, respectively) are also shown for the sake of completeness.

TABLE I
Clinical data for the acidotic and normal group. Expressed as mean (s.d.) or number (%). †marks statistical difference on $p < 0.05$.

	Acidotic n=37	Normal n=1251
Birthweight (g)	3407 (462)	3321 (469)
Operative delivery for fetal distress (n)	18 (49%)	274 (22%)†
Umbilical cord arterial pH	7.01 (0.03)	7.23 (0.07)†
Apgar score 5 minutes > 7	100%	100%
Time from end of rec. and birth (min)	2.45 (3.28)	1.68 (2.65)

TABLE II
Fetal acidosis detection clinical benchmark performance.

	SE	SP	#TP	#FP	#FN	#TN
fetal hypoxia	.49	.78	18	274	19	977

FHR time series and preprocessing. Beat to beat (RR intervals) occurrence times were extracted by the STAN devices from the recorded fetal ECGs (12-bit resolution and sampling rate 500 Hz). RR list was corrected for outliers and missing data by sliding median filtering, using a 10 second sliding window and thresholds of $(1 \pm 0.33)\bar{\mu}$, where $\bar{\mu}$ is the median computed on the window. Then, it was resampled using standard cubic spline interpolation to form regularly sampled beat-per-minute (bpm) time series $X(t)$. As FHR does not contain energy at frequencies higher than 3 Hz, the sampling frequency was set to $f_s = 10$ Hz. Missing data over durations longer than five seconds were discarded from analysis. Feature computation and supervised classification was performed on the last 20 minutes of the first stage of labor.

III. Methods: features and Sparse-SVM

A. Features

Because the present contribution focuses on sparsity in fetal acidosis detection and not on feature design, a list of 20 features was chosen, from three different categories. Enhanced and automated FIGO (FI) were selected for relevance to clinical practice; Spectral energy (SC), and multiscale analysis (SI) features were selected from previous studies [6]–[8], [11] as representative of relevant features used and tested at the research level. The complete feature set is defined as: $S_{ALL} = \{FI, SC, SI\}$.

1) FIGO enhanced and automated features (FI): Obstetricians examine FHR along three criteria: baseline evolution, decelerations/accelerations and variability. From $X(t)$, floating baseline $B(t)$ is estimated as described in [2], [12], and its evolution along time modeled by linear regression as $B(t) \approx \beta_0 + \beta_1 t$. The FIGO guidelines [2] are used here to automate the counts, $\#acc$ and $\#dec$, of accelerations and decelerations. Impact of decelerations is further quantified: i) by median absolute deviation (MAD) of $X(t)$ from

baseline: $\text{MAD}_{dtrd} = MAD(X(t) - B(t))$, quantifying the average depth ; ii) by cumulated duration of decelerations T_{stress}, quantifying average duration ; iii) by cumulated areas A_{dec} (approximated using a triangle with vertices at onset, nadir and end of deceleration), quantifying the average size. Long (LTV) and short (STV) term variabilities are computed, excluding periods of decelerations, using standard procedures, as detailed e.g., in [2], [25]. This leads to nine enhanced FIGO features automatically computed from each recording: $\{\beta_0, \beta_1, \#acc, \#dec, \text{MAD}_{dtrd}, T_{stress}, A_{dec}, LTV, STV\}$.

2) Spectral energy (SC): Spectral analysis relies on a spectral bands energy decomposition, assumed to reflect functions of the autonomic nervous system. It is chosen here to use the band definition stemming from adults [26] as no widely accepted alternative exists for fetuses (cf. [8] for discussion): very low frequency E_{VLF}: $[0.003, 0.04]$ Hz, low frequency E_{LF}: $[0.04, 0.15]$ Hz (corresponding mainly to sympathetic and parasymphatetic activity), high frequency E_{HF}: $[0.15, 0.4]$ Hz (corresponding mainly to parasymphatetic activity). Further, the ratio of these energies LF/HF, and spectral index α estimated over LF-HF bands are used [26]. Spectral estimation is conducted using a Welch periodogram (window size 1024 samples with 50% overlap). The set of spectral features (SC) is as follows: $\{E_{VLF}, E_{LF}, E_{HF}, LF/HF, \alpha\}$.

3) Multiscale multifractal analysis (SI): FHR variability is considered a key element of fetal well-being evaluation [27] and is classically quantified using LTV and STV. Following our previous contributions [7], [8], [28], variability is here quantified using the concepts of multiscale and multifractal analysis, mostly relying on wavelet transforms applied to FHR. Wavelet coefficients $d_{j,k} = \langle \psi_{j,k} | X \rangle$ are obtained by comparing FHR against a collection $\{\psi_{j,k}(t) = 2^{-j}\psi(2^{-j}t - k)\}_{(j,k)\in\mathbb{N}^2}$ of dilated (to scale 2^j) and translated (to position $2^j k$) templates of the mother wavelet ψ.They permit to measure the uniform Hölder regularity $h_{\min}$ as $\sup_k |d_{j,k}| \sim 2^{jh_{\min}}$ [29]. Further, scale-free temporal dynamics can be quantified by the so-called Hurst parameter H, measured using the evolution of energy across scales [29]: $\sum_{k=1}^{n_j} |d_{j,k}|^2 \sim 2^{2jH}$. It is now well documented that parameter H alone cannot fully account for the richness of scale-free dynamics encountered in real-world data and in heart rate variability analysis (cf. e.g., [7], [30]). This can be described in a more versatile manner by the multifractal paradigm. *Wavelet leaders* are defined as local suprema of wavelet coefficients, taken within a narrow temporal neighborhood and for all finer scales: $L_{j,k}^{(\gamma)} = \sup_{\lambda' \subset 3\lambda} 2^{j'\gamma} |d_{\lambda'}|$, with $\lambda = \lambda_{j,k} = [k2^j, (k+1)2^j)$ and $3\lambda_{j,k} = \bigcup_{m\{-1,0,1\}} \lambda_{j,k+m}$ [29]. The fractional integration parameter $\gamma \geq 0$ is chosen to ensure minimal regularity. Multifractal properties are then efficiently characterized by a vector of attributes $(c_p)_{p\geq 1}$ quantifying the evolution along scales of the statistics of log-leaders [29]: $\text{Cum}_p \ln L_{j,\cdot}^{(\gamma)} \simeq c_{p,0} + c_p \ln 2^j$. While c_1 is closely related to H [29], c_2, c_3, c_4 are associated to the evolution along scales of respectively the variance, skewness and kurtosis of $\ln L_{j,\cdot}^{(\gamma)}$, provide information beyond correlation and are thus nonlinear features. This leads to six features: $(H, h_{\min}, c_1, c_2, c_3, c_4)$.

B. Sparse-SVM (S-SVM)

1) Classical SVM: SVM has already been used for acidosis classification, cf. e.g. [15], [17]. Given N P-dimensional feature vectors $\mathbf{x}_n \in \mathbb{R}^P$ and class labels $y_n \in \{-1, 1\}$ (which correspond to healthy/non-healthy subjects), SVM searches for the optimal hyperplane $\mathbf{w}^T\mathbf{x}_n + b$, that separates (margin maximization [31]) the two classes, and obtained by solving:

$$(\widehat{\mathbf{w}}, \widehat{b}) \in \underset{\mathbf{w}\in\mathbb{R}^P,\, b\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{n=1}^{N} H_{\mathbf{w},b}(\mathbf{x}_n, y_n), \quad (1)$$

where $H_{\mathbf{w},b}(\mathbf{x}, y) = \max\big(0, 1 - y(\mathbf{w}^\top\mathbf{x} + b)\big)$ is the hinge loss function, and the regularization parameter C controls the trade-off between misclassification rate and data sparsity. Classical SVM promotes data sparsity: Only a few subjects actually contribute to determine $(\widehat{\mathbf{w}}, \widehat{b})$ at the price, though, of involving all features, i.e., $\widehat{w}_p \neq 0, \forall p$. Further, it is well documented (cf. e.g. [15]) that the curse of dimensionality, $P \approx \min(N_+, N_-)$, made particularly vivid by the nature of fetal acidosis detection (small number of acidotic cases N_+, large class imbalance ($N_+/N_- \ll 1$)), weakens the data sparsity effect of SVM, thus diminishing its classification robustness. Overcoming such difficulties has prompted research to use special feature selection algorithms (cf. e.g. [17] and approaches using SVM [32], [33]) or the use of more appealing and efficient Sparse-SVM [31], [34] that achieves feature selection and classification jointly.

2) Sparse-SVM: In a nutshell, S-SVM favors sparsity in model, i.e. in $\mathbf{w}$, rather than in data. To that end, imposing in Eq. (1) a ℓ_1-norm (instead of ℓ_2-norm) to the first term proves useful to induce sparsity in $\mathbf{w}$ [35], [36], at the price though of making the optimization problem difficult. It was however shown that the reformulation of the penalization term with a square hinge loss function, $H_{\mathbf{w},b}^2$, instead of $H_{\mathbf{w},b}$ permits to overcome this difficulty:

$$(\widehat{\mathbf{w}}, \widehat{b}) \in \underset{\mathbf{w}\in\mathbb{R}^P,\, b\in\mathbb{R}}{\operatorname{argmin}} \|\mathbf{w}\|_1 + C\sum_{n=1}^{N} H_{\mathbf{w},b}^2(\mathbf{x}_n, y_n). \quad (2)$$

To account for the imbalanced classes, the penalization term consisting of a sum $\sum_{n=1}^{N}$ across all subjects in the database is further split into two weighted sums, one for each class:

$$(\widehat{\mathbf{w}}, \widehat{b}) \in \underset{\mathbf{w}\in\mathbb{R}^P,\, b\in\mathbb{R}}{\operatorname{argmin}} \|\mathbf{w}\|_1 + C\frac{\lambda}{N_-}\sum_{n=1}^{N_-} H_{\mathbf{w},b}^2(\mathbf{x}_n, y_n) + C\frac{(1-\lambda)}{N_+}\sum_{n=1}^{N_+} H_{\mathbf{w},b}^2(\mathbf{x}_n, y_n) \quad (3)$$

with $\lambda \in (0, 1)$ an extra parameter permitting a versatile control of the balance between FP and FN detections.

C. Performance assessment

1) Performance quantification: Performance is quantified using SE, SP and the area under receiver-operation-characteristic (ROC) curve (AUC). SE and SP are evaluated as explained in Figs. 1 and 2. AUC is estimated from ROC curves, obtained by varying λ: $\lambda \to 0$, (resp. 1) favors high SE (resp., SP). For ease in comparisons (and following methodologies in [15], [19]), SP is reported for targeted SE $= 0.7$ (i.e., $\#\text{TP} = 26$, out of 37).

2) Cross validation for performance assessment: The use of supervised learning procedures, in general, and of S-SVM, in particular, requires to address three issues: i) For fixed hyper parameters (C, λ), compute decision parameters $(\widehat{\mathbf{w}}, \widehat{b})$; ii) Select optimal hyper parameters (C^{opt}, λ^{opt}) ; iii) Estimate classification performance. For the first issue, problem (3) is solved in the primal space using a Forward-Backward Splitting Algorithm, involving proximity operators [36], developed and customized by ourselves. Issues ii) and iii) are dealt with using standard cross-validation procedures [31], [37], which involve randomly and repeatedly splitting available data into training and testing subsets. *Single-loop* cross-validation (SLCV) is used to select optimal parameters C^{opt}, λ^{opt}, as detailed in Fig. 1. However, performance metrics obtained from SLCV are regarded as *optimistic* estimates, since the same dataset is used both for parameter selection and classification performance assessment [31], [37]. Relevant performance assessment requires instead the use of a *double-loop* cross-validation (DLCV) procedure, where parameters for each partition are estimated using an independent SLCV, as detailed in Fig 2. For theoretical details on the use of SLCV and DLCV, interested readers are referred to, e.g. [31], [37].

Fig. 1. Single-loop cross-validation. Feature set $\mathcal{S}$ and number of CV loops K; C and λ are hyper-parameters.

function SLCV($\mathcal{S}$, K)
for $C \in \mathcal{C}$ **do**
  for $\lambda \in \mathcal{L}$ **do**
    for $r = 1$ to R **do**
      Split $\mathcal{S}$ into disjunct sets $\{\mathcal{S}_k\}_{k=1,\dots,K}$ of size N/K.
      for $k = 1$ to K **do**
        $(\widehat{\mathbf{w}}, \widehat{b}) = \text{Learn}(\mathcal{S} \setminus \mathcal{S}_k, C, \lambda)$.
        $\hat{y}_k = \text{Predict}(\mathcal{S}_k, (\widehat{\mathbf{w}}, \widehat{b}))$.
      end /* k */
      $(SE, SP)_r = \text{PerfEval}(y, \{\hat{y}_k\}_{k=1,\dots,K})$
    end /* r */
    $(SE, SP)_{C,\lambda} =< (SE, SP)_r >_r$
  end /* λ */
  $\lambda_C = \max_\lambda\{(SE, SP)_{C,\lambda} | SE \geq 0.7\}$.
  $\widehat{\mathbf{w}}_C = \text{Learn}(\mathcal{S}, C, \lambda_C)$.
end /* C */
$(C^{opt}, \lambda^{opt}) = \max_{C,\lambda}\{(SE, SP)_{C,\lambda} | SE \geq 0.7\}$.
$(\mathbf{w}^{opt}, b^{opt}) = \text{Learn}(\mathcal{S}, C^{opt}, \lambda^{opt})$.
$AUC_{C^{opt}} = \text{ROC}((SE, SP)_{C^{opt},\lambda})$.
output $\{C^{opt}, \lambda^{opt}, (SE, SP)_{C^{opt},\lambda^{opt}}, AUC_{C^{opt}}\}$

Fig. 2. Double-loop cross-validation. Feature set $\mathcal{S}$, number of outer CV loops K, and number of inner CV loops J; C and λ are hyper-parameters.

function DLCV($\mathcal{S}$, K, J)
for $r = 1$ to R **do**
  Split $\mathcal{S}$ into disjunct sets $\{\mathcal{S}_k\}_{k=1,\dots,K}$ of size N/K.
  for $k = 1$ to K **do**
    $(C_{r,k}, \lambda_{r,k}) = \text{SLCV}(\mathcal{S} \setminus \mathcal{S}_k, J)$
    $(\widehat{\mathbf{w}}, \widehat{b}) = \text{Learn}(\mathcal{S} \setminus \mathcal{S}_k, C_{r,k}, \lambda_{r,k})$.
    $\hat{y}_k = \text{Predict}(\mathcal{S}_k, (\widehat{\mathbf{w}}, \widehat{b}))$.
  end /* k */
  $(SE, SP)_{r,k} = \text{PerfEval}(y, \{\hat{y}_k\}_{k=1,\dots,K})$
end /* r */
$(SE, SP) =< (SE, SP)_{r,k} >_{r,k}$
$(C^{opt}, \lambda^{opt}) = \text{SLCV}(\mathcal{S}, K, C, \lambda)$
$(\mathbf{w}^{opt}, b^{opt}) = \text{Learn}(\mathcal{S}, C^{opt}, \lambda^{opt})$.
output $\{< C_{r,k} >_{r,k}, < \lambda_{r,k} >_{r,k}, SE, SP\}$

Number of CV folds. A leave-one-out CV is used for performance assessment: The numbers of CV loops are set to $K = N_+$ (for the CV outer loop of DLCV) and to $J = N_+ - 1$ (for the inner loop SLCV). This procedure provides robust parameter estimation $(\mathbf{w}, C, \lambda)$ and straightforward analysis of misclassified acidotic cases. Further, because CV splits randomly data into training and testing, the CV is repeated $R = 10$ times, for each acidotic, to minimize the impact of the random selection in healthy subjects.

Feature preprocessing. Because the values within which each feature naturally evolves are very different, features are first preprocessed for outliers (Winsorization in the interval $[Q_1 - 3IQR, Q_3 + 3IQR]$, where Q_i is the i-th quartile and $IQR = Q_3 - Q_1$) and standardized (zero mean and unit variance) across the entire population N.

IV. Results

A. Univariate analysis

For benchmark, each feature is first used alone for fetal acidosis detection, in a SLCV S-SVM procedure. Table III shows that, individually, features A_{dec}, MAD_{dtrd}, T_{stress}, E_{VLF}, H, and c_1 yield the best performance. They are slightly better than those obtained in clinical practice (cf. Table II), with SP $\approx 0.60 - 0.65$ for the targeted SE $= 0.7$. These features clearly point to the importance of decelerations (duration and depth, but not their numbers) and variability in acidosis detection. Further, Table III indicates that scale-free variability, as measured by H (or c_1), outperforms traditional STV and LTV, which favor a priori selected specific time scales. Nonlinear scale free-features $\{c_2, c_3, c_4\}$ do not have significant individual power, neither do the linear spectral index α and LF/HF ratio (corresponding to results in [8]).

Despite satisfactory SE, the low SP that is achieved prompts for the joint use of multiple features in acidosis detection. Prior to multivariate classification and to help interpretation, correlations amongst all pairs of features, reported in Fig. 3, are analyzed: i) MAD_{dtrd}, A_{dec}, T_{stress} and E_{VLF} are correlated, indicating that E_{VLF} are associated with decelerations ; ii) Baseline features do not correlate with other features ; iii) Features $H, c_1, h_{\min}$ are found highly correlated, which comes as no surprise given their definitions [29] ; iv) Nonlinear features $\{c_p\}_{p\geq 2}$ are weakly correlated to linear features ; v) As expected, energies in frequency bands and LF/HF ratio are correlated [8].

TABLE III
Univariate performance. SP and #FP correspond to targeted SE=0.70, #TP=26, computed using SLCV.

Feature:	β_0	β_1	$\#acc$	$\#dec$	A_{dec}	MAD_{dtrd}	T_{stress}	STV	LTV	LF/HF
AUC	.65	.51	.52	.57	.69	.73	.70	.50	.50	.53
SP	.58	.22	.19	.40	.64	.63	.65	.27	.27	.39
#FP	525	980	1015	748	449	462	436	912	913	768
Feature:	E_{VLF}	E_{LF}	E_{HF}	α	H	$h_{\min}$	c_1	c_2	c_3	c_4
AUC	.71	.56	.52	.58	.71	.68	.70	.61	.50	.52
SP	.65	.30	.30	.43	.60	.52	.54	.38	.27	.20
#FP	440	877	878	712	500	606	575	775	914	998

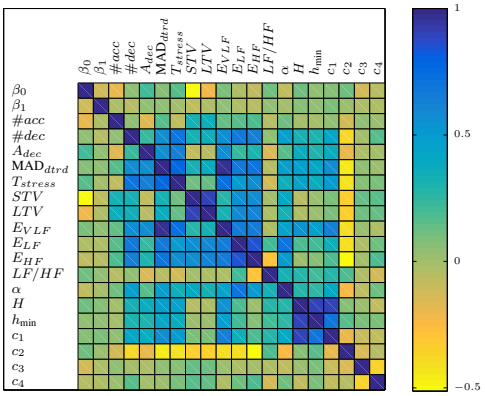

Fig. 3. Pair-wise correlation matrix of all features.

B. Multivariate analysis

1) Feature selection: Fig. 4 reports the weights, obtained from SLCV, for each feature as function of C and shows: i) E_{VLF} is marginally chosen at very low C but is then excluded ; ii) For correlated features, such as (H, c_1) or $\text{MAD}_{dtrd}, T_{stress}, A_{dec}, E_{VLF})$, only one is selected ; iii) MAD_{dtrd}, β_0 and H are the three first features selected and they remain included while increasing C ; iv) Nonlinear features c_2 and c_4 are included next.

Fig. 5 reports AUC and SP (from SLCV, cf. Fig 1), as functions of C. It shows that optimal performance is obtained for high sparsity ($C = 2^{-8.5}$): Only three features (out of 20) are combined to yield optimal acidosis detection. The selected features $(\text{MAD}_{dtrd}, \beta_0, H)$ are interestingly observed to correspond to categories of FHR properties used in clinical practice: β_0 quantifies the average level of the baseline ; MAD_{dtrd} quantifies the overall impact of decelerations ; H provides a robust measure of FHR variability.

The same SLCV procedure is applied to the three sub-groups (FI, SC, SI) of features independently. Interestingly, it is observed (cf. Table IV) that while MAD_{dtrd} and β_0 are selected first, FIGO features alone would not yield a sparse representation, involving six out of the nine features. This shows that when added to FIGO features, H efficiently replaces a number of them. Further, when used alone, the SI features do take advantage of the nonlinear features c_2 and c_4. Also, when used alone, the SC features show the role of E_{VLF} (decelerations) and E_{HF} (variability).

2) Performance assessment: Again for pedagogical purposes, Table V shows performance evaluated using the SLCV (bottom) and DLCV (top) procedures. Let us note that with DLCV the λ can no longer be varied and the targeted SE of $\#\text{TP} = 26$ cannot be controlled, explaining the mild variations observed in the #TP column. The AUC cannot be computed for the same reason. Also the optimal C from DLCV does not necessarily converge to those C due to the data variations in different CV loops. The table indicates that SLCV performance is optimistically biased compared to that more accurately estimated from DLCV, both when analyzing all 20 features and when applied to each subgroup independently. More importantly, Table V shows that the triplet of features $(\text{MAD}_{dtrd}, \beta_0, H)$, extracted directly from the full set of 20 features, displays better performance than any subgroup (cf. Table V). Compared to the best univariate performance $\text{SP} = 0.65$ ($\#\text{FP} = 436$), it is observed that for the targeted $\text{SE} = 0.7$ ($\#\text{TP} = 26$), SP is significantly increased up to $\text{SP} = 0.75$ ($\#\text{FP} = 317$). Also, for this triplet performance degrades only mildly when going from SLCV to DLCV procedures, satisfactorily indicating robust and generalizable ability to perform acidosis detection on unseen data. Further, such performance satisfactorily outperforms those obtained in clinical practice (cf. Table II).

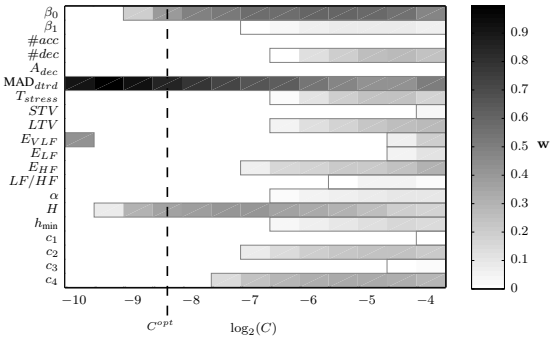

Fig. 4. Weights $\mathbf{w}$ as functions of C. Small C forces $\mathbf{w}$ to be sparse, e.g. for $\log_2(C) = -10$ only MAD_{dtrd} and E_{VLF} are selected.

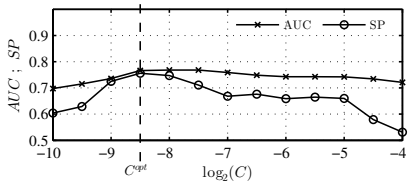

Fig. 5. Performance (SP, AUC) for $\text{SE} \geq .7$ in SLCV as functions of C.

TABLE IV
SELECTED FEATURES. WEIGHTS $\mathbf{w}$ ARE COMPUTED USING SLCV.

S_{ALL}	$\mathbf{w}$	FI	$\mathbf{w}$	SC	$\mathbf{w}$	SI	$\mathbf{w}$
MAD_{dtrd}	.84	MAD_{dtrd}	.78	E_{VLF}	.96	H	.87
β_0	.39	β_0	.58	E_{HF}	.28	c_2	.37
H	.36	T_{stress}	.16			c_4	.34
		#dec	.14				
		β_1	.08				
		LTV	.04				

TABLE V
SPARSE-SVM PERFORMANCE. TOP: DLCV, BOTTOM: SLCV .

	$\log_2 C$	AUC	SE	SP	#TP	#FP
S_{ALL}	-8.5	–	.73	.75	27	317
FI	-6	–	.70	.72	26	351
SC	-5	–	.68	.64	25	448
SI	-6	–	.65	.58	24	529
S_{ALL}	-8.5	.77	.70	.76	26	305
FI	-6	.75	.70	.74	26	330
SC	-4.5	.66	.70	.67	26	418
SI	-6	.73	.70	.63	26	458

V. Discussion

A. Confusion table for pH and selected features

Beyond quantitative performance assessment, let us comment on the quality of the achieved classification by examination of the characteristics of its confusion table, i.e., of the four classification groups: TP, FP, FN, TN. Fig. 6 displays boxplots quantifying the distributions of pH for these four groups. A Wilcoxon rank sum test concludes no difference ($pval > 0.05$) between correctly (TP) or incorrectly (FN) classified acidotic subjects. This indicates that TP and FN subjects have a similarly low pH, i.e. are equally unhealthy. Conversely, there is a statistically significant difference between correctly (TN) and incorrectly (FP) classified healthy subjects ($pval \approx 10^{-16}$). This indicates that pH for the FP varies in a lower range, compared to the *normal* range within which TN's pH are typically observed to evolve (cf. e.g. [38]). In short, FP subjects are *less healthy* than TN, and it may be considered that their pH has started to degrade.

Figs. 6 and 7 further analyze the characteristics of the four groups with respect to the (projected) decision boundary using the three features $(\text{MAD}_{dtrd}, \beta_0, H)$ contributing to classification. It shows, as expected, that TP and FP groups are very similar for all three features. The same holds for TN and FN groups, as well as for other features. This shows that using these three features, or any other in the list of 20, FP cannot be discriminated from TP and that FN can not be distinguished from TN. This is further confirmed in Fig. 8, displaying FHR time series, that can be considered as typical for each group: TP and FP records have higher baseline, larger and more frequent decelerations, and lower variability. Subjects from the FP group thus exhibit typical patterns leading to fetal hypoxia, such as large and frequent decelerations and reduced variability, that might however have lasted only over short durations without developing fetal acidosis. This detailed case study prompts for a continued search for new and more discriminative features.

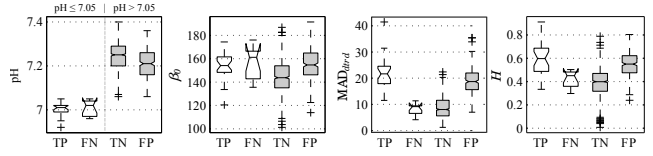

Fig. 6. Distribution of pH, β_0, MAD_{dtrd}, and H for TP, FN, TN, and FP groups.

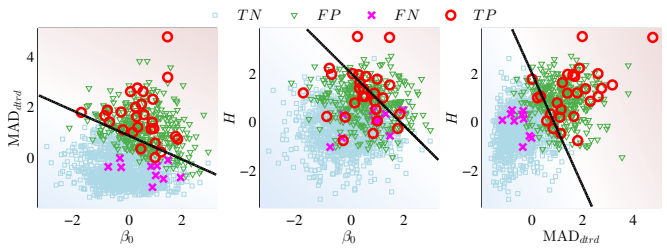

Fig. 7. Feature space $(\text{MAD}_{dtrd}, \beta_0, H)$ projections with decision boundary.

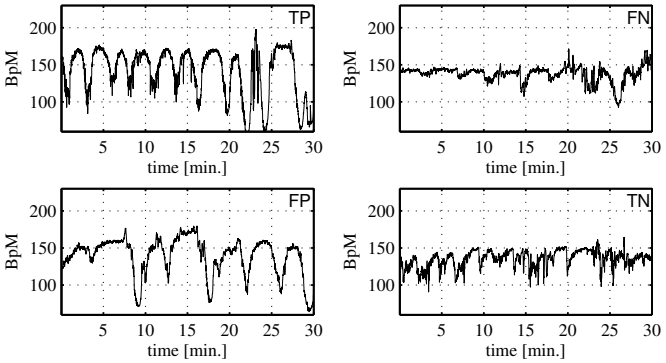

Fig. 8. Typical FHR patterns for the four groups.

B. Boolean combination of feature-group

The 20 features we chose to use stemmed from three different classes: FI, SC, SI. In the results reported above, it has been chosen to feed S-SVM with the concatenated lists of features. However, other approaches have been investigated [15], [19]. S-SVM was first applied to each group independently, with performances as reported in Table V. Second, it was investigated whether boolean combination of the decision achieved by pairs of groups (e.g., FI and SI) could yield better performance. Results, reported in Table VI, indicate that SLCV performance was slightly better while DLCV performance degraded significantly, showing a weak generalization ability due to overfitting. First, this highlights the importance of DLCV for correct performance estimation, which must be independent of hyper-parameter selection. Second, it indicates that more *complex* models (in number of features and parameters) are less robust than *simple* ones.

TABLE VI
Conjunction of feature groups. Top: DLCV, Bottom: SLCV.

$G_1 \wedge G_2$	$\log_2 C_1$	$\log_2 C_2$	SE	SP	#TP	#FP
$FI \wedge SC$	-7.5	-7.0	.62	.74	23	328
$FI \wedge SI$	-7	-6.0	.60	.76	22	307
$SC \wedge SI$	-6.5	-5.5	.68	.69	25	389
$FI \wedge SC$	-8	-4.5	.70	.76	26	297
$FI \wedge SI$	-8	-5.5	.70	.77	26	288
$SC \wedge SI$	-8.5	-5.5	.70	.72	26	351

C. Relevance of features and performance assessment

Standalone (or univariate) use of features for acidosis detection permits to show that some features achieve by themselves detection performances that are (slightly) better than the clinical benchmark reported in Table II. Interestingly, such features essentially consist of enhanced and automatized versions of FIGO criteria (A_{dec}, MAD_{dtrd}, T_{stress}), rather than products of advanced signal processing tools (though E_{VLF}, H, and c_1 also perform well). Obtained results confirm earlier reports on several different datasets (cf. e.g., [8], [16], [18]) indicating that scale-free measures of variability (H and c_1) provide valuable and relevant information for acidosis detection. In effect, Table III shows that STV and LTV measures of variability, FIGO-based and tied to specific time

scales, perform less well than scale-free ones (H and c_1), which involve a continuum of time scales (see also [8]).

Multivariate supervised classification, designed from a S-SVM machine learning strategy, was then shown to permit both to select a sparse and significant subset of features and to achieve optimal acidosis detection performance. While S-SVM never selects a feature that performs poorly individually, it removes features that individually perform well because they are redundant (correlated). Interestingly, when examining pairs of highly correlated features, such as MAD_{dtrd} and E_{VLF}, it can be observed that it is not necessarily the one that performs best that is selected by S-SVM, thus underlining the benefits of S-SVM regularization in the multivariate procedure. Further, it was found, from the large database used here, that S-SVM preferred a high level of sparsity, with only three features combined to achieve optimal fetal acidosis detection, yielding a low complexity decision rule, and hence favoring its interpretation and actual use by obstetricians. Notably, these three features, $\mathrm{MAD}_{dtrd}, \beta_0$ and H, match the FIGO guidelines advising the joint examination of baseline, decelerations and variability to assess FHR signals. Moreover, both univariate and multivariate classification emphasize the predominance of decelerations in fetal acidosis detection, in agreement with documented medical reports on the relations of deceleration to fetus stress and oxygen supply [27] and earlier reports on acidosis detection performance, cf. e.g., [39]. The performance obtained from the sparse and simple model constructed by S-SVM ($\mathrm{SE} = 0.73$, $\mathrm{SP} = 0.75$), significantly outperforms both univariate performance of the best standalone feature, and the clinical benchmark performance. Finally, the presented work contributes to the state of the art in several aspects: i) S-SVM model is trained and tested on significantly larger database than is typically reported in literature, cf. e.g., [15], [17]–[19] ii) S-SVM model is not only sparse, allowing easy interpretation and transfer into clinical practice, but also achieves results that outperform those obtained with different tools on different data sets (cf. e.g., [15], [17], [19]). Further, performance was estimated using a double-loop cross validation strategy (as opposed to single-loop cross validation), which was shown to be a critical step in correctly assessing actual detection performance on unseen data, i.e., having robust generalization properties. This comes at the cost of significantly increased computational burden but confirms that simple models generalize better than complex ones. To avoid such issues, independent databases may be used for (hyper)-parameter tuning and for performance assessment. The preliminary results with the database described here and the open access CTG database [40] are reported in [41].

Two key remarks are further in order: First, the presented work uses $\mathrm{pH} \leq 7.05$ as a marker of fetal acidemia. Even-though this value is not consensual it is widely used in other studies, cf. e.q. [6], [10], [14], [17], [24] and well accepted in general clinical practice. Second, it was chosen here to maintain the population ratio of normal and acidotic cases, and not to artificially subsample the large class of healthy cases as it was performed in other studies, cf. e.g. [15]–[17]. This allows to reproduce in the database the prevalence of acidotic cases ($\sim 3\%$ corresponds to general practice [42]), and to make the analysis independent of subsampling strategies.

VI. Conclusions

This contribution aims to promote Sparse Support Vector Machine classification, that allows both to select (a small number of) relevant features within a large list and to achieve efficient fetal acidosis detection. It shows that the automatic selection of a sparse subset of features achieves satisfactory classification performance (sensitivity 0.73 and specificity 0.75, outperforming clinical practice). This contribution thus improves intrapartum fetal acidosis detection in several respects: A set of features combining FIGO, spectral and scale-free dynamics attributes is used ; an original multivariate classification targeting both sparse feature selection and high performance is devised ; state-of-the-art performance is obtained on a much larger database than those generally studied.

Despite satisfactory and promising classification performance, the detailed case study of the classification confusion table indicates that there is still room for improvements, as the three selected features show a weak power to further discriminate the remaining FPs from the TPs, and the FNs from the TNs. Improvements may come from two directions. First, new features could be used. The set of features used here may have been longer or different, notably the nonlinear *entropy rates* [9], widely used in the literature, *phase-rectified signal average* [13], or the more recently proposed *scattering features* [10], were observed not to improve detection performance. This means that the search for new and original characterization of FHR must remain an open and active field, likely developing at the interface between obstetricians and signal processing experts to incorporate medical knowledge in mathematical tools. Second, the present contribution focuses on a static window of 20 minutes at the very end of the first stage of labor (following the analysis in [19], [23]). However, preliminary attempts, cf. e.g. [10], [15], [43] indicate that further improvement in performance may be gained from analyzing the evolution of FHR features along time. This is being further investigated.

References

[1] E. Chandraharan and S. Arulkumaran, "Prevention of birth asphyxia: responding appropriately to cardiotocograph (CTG) traces." *Best Pract Res Clin Obstet Gynaecol*, vol. 21, no. 4, pp. 609–624, Aug 2007.

[2] FIGO, "Guidelines for the Use of Fetal Monitoring," *Int J Gynaecol Obstet*, vol. 25, pp. 159–167, 1986.

[3] S. C. Blackwell *et al.*, "Interobserver and intraobserver reliability of the NICHD 3-Tier Fetal Heart Rate Interpretation System." *Am J Obstet Gynecol*, vol. 205, no. 4, pp. 378.e1–378.e5, Oct 2011.

[4] Z. Alfirevic *et al.*, "Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour." *Cochrane Database Syst Rev*, vol. 3, no. 3, p. CD006066, 2006.

[5] J. T. Parer *et al.*, "Fetal acidemia and electronic fetal heart rate patterns: is there evidence of an association?" *J Matern Fetal Neonatal Med*, vol. 19, no. 5, pp. 289–294, May 2006.

[6] S. M. Siira *et al.*, "Marked fetal acidosis and specific changes in power spectrum analysis of fetal heart rate variability recorded during the last hour of labour." *BJOG*, vol. 112, no. 4, pp. 418–423, Apr 2005.

[7] M. Doret *et al.*, "Multifractal analysis of fetal heart rate variability in fetuses with and without severe acidosis during labor," *Am J Perinatol*, vol. 28, no. 4, p. 259, 2011.

[8] ——, "Fractal Analysis and Hurst Parameter for intrapartum fetal heart rate variability analysis: A versatile alternative to Frequency bands and LF/HF ratio." *PLoS ONE*, vol. 10, no. 8, p. e0136661, 08 2015.

[9] M. Costa *et al.*, "Multiscale entropy analysis of biological signals," *Phys Rev E*, vol. 71, no. 2 Pt 1, p. 021906, Feb 2005.

[10] V. Chudáček *et al.*, "Scattering Transform for Intrapartum Fetal Heart Rate Variability Fractal Analysis: A Case-Control Study," *IEEE Trans Biomed Eng*, vol. 61, no. 4, pp. 1100–1108, April 2014.

[11] H. Gonçalves *et al.*, "Linear and nonlinear analysis of heart rate patterns associated with fetal behavioral states in the antepartum period," *Early Hum Dev*, vol. 83, no. 9, pp. 585–591, Sep 2007.

[12] V. Chudáček *et al.*, "Automatic evaluation of intrapartum fetal heart rate recordings: A comprehensive analysis of useful features," *Physiol meas*, vol. 32, pp. 1347–1360, 2011.

[13] A. Georgieva *et al.*, "Phase-rectified signal averaging for intrapartum electronic fetal heart rate monitoring is related to acidaemia at birth." *BJOG*, vol. 121, no. 7, pp. 889–894, Jun 2014.

[14] A. Costa *et al.*, "Prediction of neonatal acidemia by computer analysis of fetal heart rate and ST event signals." *Am J Obstet Gynecol*, vol. 201, no. 5, pp. 464.e1–464.e6, Nov 2009.

[15] P. Warrick *et al.*, "Classification of normal and hypoxic fetuses from systems modeling of intrapartum cardiotocography," *IEEE Trans Biomed Eng*, vol. 57, no. 4, pp. 771–779, 2010.

[16] J. Spilka *et al.*, "Using nonlinear features for fetal heart rate classification," *Biomed Signal Proces*, vol. 7, no. 4, pp. 350–357, 2012.

[17] L. Xu *et al.*, "Feature selection using genetic algorithms for fetal heart rate analysis," *Physiol meas*, vol. 35, no. 7, pp. 1357–71, 2014.

[18] M. G. Frasch *et al.*, "Correlating multidimensional fetal heart rate variability analysis with acid-base balance at birth," *Physiological Measurement*, vol. 35, no. 12, p. L1, 2014.

[19] S. Dash *et al.*, "Fetal Heart Rate Classification Using Generative Models," *IEEE Trans Biomed Eng*, vol. 61, no. 11, pp. 2796–2805, Nov 2014.

[20] R. Czabanski *et al.*, "Computerized analysis of fetal heart rate signals as the predictor of neonatal acidemia," *Expert Systems with Applications*, vol. 39, no. 15, pp. 11 846–11 860, 2012.

[21] A. Georgieva *et al.*, "Artificial neural networks applied to fetal monitoring in labour." *Neural Computing and Applications*, vol. 22, no. 1, pp. 85–93, 2013.

[22] 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*, vol. 156, no. 1, pp. 35–40, May 2011.

[23] 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*, 2014, pp. 777–780.

[24] 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*, vol. 358, no. 9281, pp. 534–538, Aug 2001.

[25] J. Pardey *et al.*, "A computer system for the numerical analysis of nonstress tests." *Am J Obstet Gynecol*, vol. 186, no. 5, pp. 1095–1103, May 2002.

[26] Task Force of the European Society of Cardiology and the American Society of Pacing and Electrophysiology, "Heart rate variability: standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, p. 104365, 1996.

[27] A. Ugwumadu, "Understanding cardiotocographic patterns associated with intrapartum fetal hypoxia and neurologic injury." *Best Pract Res Clin Obstet Gynaecol*, vol. 27, no. 4, pp. 509–536, Aug 2013.

[28] R. Leonarduzzi *et al.*, "p-leader multifractal analysis and sparse SVM for intrapartum fetal acidosis detection," in *Ann Int Conf IEEE Eng Med Biol Soc (EMBC)*. Milan, Italy, Aug. 2015, pp. 1971 – 1974.

[29] H. Wendt, P. Abry, and S. Jaffard, "Bootstrap for Empirical Multifractal Analysis," *IEEE Signal Proc. Mag.*, vol. 24, no. 4, pp. 38–48, 2007.

[30] P. Ivanov *et al.*, "Multifractality in human heartbeat dynamics," *Nature*, vol. 399, no. 6735, pp. 461–465, 1999.

[31] T. Hastie *et al.*, *The elements of statistical learning: data mining, inference and prediction*, 2nd ed. Springer, 2009.

[32] E. Bron *et al.*, "Feature selection based on the svm weight vector for classification of dementia," *Biomedical and Health Informatics, IEEE Journal of*, vol. 19, no. 5, pp. 1617–1626, Sept 2015.

[33] C. Soguero-Ruiz *et al.*, "Support vector feature selection for early detection of anastomosis leakage from bag-of-words in electronic health records," *Biomedical and Health Informatics, IEEE Journal of*, vol. PP, no. 99, pp. 1–1, 2015.

[34] L. Laporte *et al.*, "Nonconvex Regularizations for Feature Selection in Ranking With Sparse SVM," *IEEE Trans Neural Netw Learn Syst*, vol. 25, no. 6, pp. 1118–1130, 2014.

[35] M. Blondel *et al.*, "Block coordinate descent algorithm for large-scale sparse multiclass classification," *J Mach Learn*, vol. 93, no. 1, pp. 31–52, Oct. 2013.

[36] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Model Simul*, vol. 4, no. 4, pp. 1168–1200, 2005.

[37] P. Jonathan *et al.*, "On the use of cross-validation to assess performance in multivariate prediction," *Stat Comput*, vol. 10, no. 3, pp. 209–229, 2000.

[38] A. Georgieva *et al.*, "Umbilical cord gases in relation to the neonatal condition: the EveREst plot," *Eur J Obstet Gyn R B*, vol. 168, no. 2, pp. 155 – 160, 2013.

[39] J. A. Westgate *et al.*, "The intrapartum deceleration in center stage: a physiologic approach to the interpretation of fetal heart rate changes in labor." *Am J Obstet Gynecol*, vol. 197, no. 3, pp. 236.e1–236.11, Sep 2007.

[40] V. Chudáček *et al.*, "Open access intrapartum CTG database." *BMC Pregnancy Childbirth*, vol. 14, no. 1, p. 16, 2014.

[41] J. Spilka *et al.*, *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016: MEDICON 2016, April 2016, Cyprus*. Springer International Publishing, 2016. Accepted., ch. Intrapartum Fetal Heart Rate Classification: Cross-Database Evaluation.

[42] P. Yeh *et al.*, "The relationship between umbilical cord arterial ph and serious adverse neonatal outcome: analysis of 51,519 consecutive validated samples." *BJOG*, vol. 119, no. 7, pp. 824–831, Jun 2012.

[43] J. Spilka *et al.*, "Intrapartum Fetal Heart Rate Classification from Trajectory in Sparse SVM feature space," in *Ann Int Conf IEEE Eng Med Biol Soc (EMBC)*. Milan, Italy, August 2015, pp. 2335–2338.