# Multi-feature classifiers for burst detection in single EEG channels from preterm infants

M Navarro, M Porée, M Kuchenbuch, M Chavez, Alain Beuchée, G Carrault

# Multi-feature classifiers for burst detection in single EEG channels from preterm infants

X. Navarro[1]*, F. Porée[2,3] , M. Kuchenbuch[2,3,4] , M. Chavez[5] , Alain Beuchée[2,3,4] and G. Carrault[2,3]

[1]Sorbonne Universités, UPMC Univ Paris 06, INSERM UMRS-1158 Neurophysiologie Respiratoire Expérimentale et Clinique, Paris, France

[2]Université de Rennes 1, Laboratoire Traitement du Signal et de l'Image, Rennes, France
[3]INSERM U1099, Rennes, France
[4]CHU Rennes, Pôle de pédiatrie médico-chirurgicale et de génétique clinique, Rennes, France
[5]CNRS UMR7225, Hôpital de la Pitié Salpêtrière Paris, France

## Abstract

*Objective:* The study of electroencephalographic (EEG) bursts in preterm infants provides valuable information about maturation or prognostication after perinatal asphyxia. Over the last two decades, a number of works proposed algorithms to automatically detect EEG bursts in preterm infants, but they were designed for populations under 35 weeks of post menstrual age (PMA). However, as the brain activity evolves rapidly during postnatal life, these solutions might be under-performing with increasing PMA. In this work we focused on preterm infants reaching term ages (PMA $\geq$ 36 weeks) using multi-feature classification on a single EEG channel. *Approach:* Five EEG burst detectors relying on different machine learning approaches were compared: Logistic regression (LR), linear discriminant analysis (LDA), k-nearest neighbors (kNN), support vector machines (SVM) and thresholding (Th). Classifiers were trained by visually labeled EEG recordings from 14 very preterm infants (born after 28 weeks of gestation) with 36 – 41 weeks PMA. *Main results:* The most performing classifiers reached about 95% accuracy (kNN, SVM and LR) whereas Th obtained 84%. Compared to human-automatic agreements, LR provided the highest scores (Cohen's kappa = 0.71) using only three EEG features. Applying this classifier in an unlabeled database of 21 infants $\geq$ 36 weeks PMA, we found that long EEG bursts and short inter-burst periods are characteristic of infants with the highest PMA and weights. *Significance:* In view of these results, LR-based burst detection could be a suitable tool to study maturation in monitoring or portable devices using a single EEG channel.

***Index Terms*— EEG bursts, preterm infants, automated detection, logistic regression**

[1]Author for correspondence: xavier.navarro@upmc.fr

## 1 Introduction

The electroencephalographic (EEG) activity in preterm infants is characterized by discontinuous patterns, alternating quiescence periods with slow, high voltage transients or bursts, continuously evolving during infancy. Inter-burst intervals (IBIs) provide valuable information about progostation and maturation as they progressively decrease in duration with increasing age in healthy preterm infants [1,2]. As preterm infants reach term age (37 to 40 PMA) EEG becomes more complex, but immature patterns are still present [3, 4]. At this stage, a *tracé alternant* (an alternating pattern of bursts and relatively quiet periods [5]) predominates in quiet sleep, but long IBIs may appear, suggesting an abnormal neurodevelopmental outcome [6]. Therefore, studying IBIs and transients in the EEG after term equivalent ages can be useful as an early prognostic tool [7].

Nevertheless, the study of IBIs is not a clinical routine in neonatal intensive care units (NICUs) and, to our knowledge, it is not exploited in portable and home monitoring devices. In effect, NICUs prioritize other vital signs over electroencephalography, which is tedious and time-consuming. Moreover, the manual recognition of IBIs and bursts is, still, another laborious and subjective task requiring trained neurologists. The automation of this procedure would therefore save time costs, avoid disagreement between different annotators and, in turn, gain attractiveness as a monitoring tool in the NICU. This challenge has motivated a number of works that propose different approaches, including supervised learning (single or multi-feature based) and clustering. The choice of the most appropriate solution is not always straightforward and relies on both clinical (e.g. infants age, developmental problems) and technical criteria (e.g number of available channels).

The vast majority of existing algorithms have been designed for burst suppression applications both in adults (anesthesia or coma monitoring) and full-term newborns

(EEG monitoring after perinatal asphyxia), but only a few works proposed burst detectors for preterm EEG. Even if similarities exist, burst suppression patterns are related to a clinical condition whereas preterm's bursts describe the normal EEG and evolve during postnatal life. In Table 1, we provide a summary of the works published over the last two decades related to preterm burst detection as well as some relevant solutions for burst suppression in full-term infants and adults.

If the brain activity needs to be segmented into burst and IBIs, binary classification or regression can be employed. Althoug this constitutes the norm of burst classification, in certain cases a third class (artefacts [9], continuous pattern [14]) or other categories (classification of different degrees of activity after asphysia [15,19]) can be considered.

In burst detection, single-feature detectors are often preferred when the EEG patterns are predominantly dichotomous, whose low frequency deflections from the baseline allow the use of direct measures (such as voltage amplitude) or functions applied on the EEG (such as energy) as only feature. Dichotomous patterns can be found in a variety of altered states of consciousness [22], but in healthy preterm infants they are characteristic of ages below 32 weeks PMA [23]. Single-feature detectors are fast and can be easily implemented by simple thresholding. Thresholding has been successfully employed for burst detection from a single EEG channel in very preterm infants [24] using the Teager-Kaiser or nonlinear energy operator (NLEO) [25]. NLEO and its variants [18,26] are widely employed for EEG burst detection (see Table 1). For multichannel data, thresholding can also be applied in successive steps using EEG power [14], NLEO [20] or using line length as feature [16].

In general, using a single feature performs fairly well for very immature patterns, but as EEG complexity increases, supplementary descriptors are needed. Thus, detecting EEG patterns in full-term newbors often requires the use of several features, as those derived from wavelet analysis for *tracé alternant* detection [8] or a variety of time and frequency-based descriptors for burst suppression detectors [11, 13].

In this work, we address the detection of bursts in very preterm infants who reached term-equivalent ages (TEA). Considering the growing field of portable EEG headsets and wearable sensors, we studied the viability of burst detection using a single EEG channel. Logistic regression (LR) is evaluated by using experts visual marks and compared to other popular multi-feature methods such as linear discriminant analysis (LDA), SVM, k-nearest neighbor (kNN) and the most commonly used single-feature classifier: Thresholding (Th).

The remainder of the paper is as follows: Section 2 describes the database and the procedure to build the reference labels. In Section 3, we present the employed classifiers and the evaluation methodology. Section 4 compares automatic and visual detections and shows the results of applying the LR based classifier to assess the maturation in our cohort. A discussion, including the strengths and limitations of this study, is provided in Section 5 and some concluding remarks are finally drawn in Section 6.

# 2 Database

## 2.1 EEG recordings

Thirty-one very preterm infants born after 27 to 29 weeks of gestation, were recorded at the CHU Hospital at Rennes (France) to study the effects of immunization (see [27] for more details about the protocol). Only pre-immunization recordings were considered in the present work to avoid eventual perturbations following the administration of vaccine. Infants, who presented a normal outcome and had discharged home, accounted for at least seven weeks of postnatal life (36 to 41 weeks PMA) during the recordings. The study was approved by the local institutional ethics committee (*Comité de Protection des Personnes*, CPP Ouest 6-598, France) and a written informed consent was given by parents.

For each newborn, two EEG channels were acquired at sampling frequency $F_s$=512 Hz using a Brainz© bedside monitor (Natus Medical Incorporated, San Carlos, USA) during 2 to 3 hours. Hydrogel surface electrodes were placed in fronto-parietal and temporal positions, corresponding approximately to the Fp1, Fp2, T3 and T4 locations of 10-20 standard systems. A bipolar reference was applied to obtain two channels (Fp1-T3, Fp2-T4). Additionally, electrocardiogram (ECG), respiratory activity and hypnograms (annotations of the sleep stages in conformity to the neonatal standard [28]) were available.

## 2.2 Annotated dataset

The annotated dataset provided the ensemble of EEG segments visually evaluated by two experienced neaonatologists, who independently marked this dataset in burst/IBI periods. Visual evaluations were available for $N_e$=14 segments from the main database of 31 infants. Segments, of length $D = 300$ seconds, were obtained from 14 different infants (36.1 to 39.7 weeks PMA). To ensure the existence of discontinuous or semi-discontinuous patterns, only segments free of artifacts in quiet sleep were considered.

| Description | Patients | #Chan | #Sc | Features | Classifier | Performance |
|---|---|---|---|---|---|---|
| Detection of tracé alternant during sleep [8] | 6 full-term | 14 | 1 | Discrete wavelet transform | S; Th | n/a |
| Burst suppression during anesthesia [9] | 17 adults | 1 | 1 | 1 (Nonlinear energy operator) | S; Th | Acc=94% |
| EEG bursts & heart beat ratio relationship [10] | 15 full-term | 1 | 1 | 1 (Instant. variance) | U; Th | n/a |
| Burst suppression detection after asphyxia [11] | 6 full-term | 8 | 1 | 5 (Energy and frequency based) | S; SVM | AUC=0.96 |
| Burst detection in extremely preterm [12] | 18 preterm (23-28/28-30 w. PMA) | 1 | 2 | 1 (Nonlinear energy operator) | S; Th | Acc=90/81% |
| Burst suppression detection [13] | 26 full-term | 8 | 1 | 9 (Energy and frequency based) | S; FLD | Acc=94% |
| Burst, IBI and continuous EEG detection [14] | 8 early preterm (29-34 w PMA) | 18 | 2 | 1 (EEG power in multiple channels) | S; Th | Sn=90% (Bursts) Sn=80% (IBIs) |
| IBI adaptive segmentation in encephalopathy [15] | 8 full-term | 13 | 1 | 1 (Amplitude) | S; Th | n/a |
| Burst detection in preterms [16] | 13 preterm (26-34 w PMA) | 9 | 2 | 1 (Line length) | S; Th | Acc=84% |
| Burst detection & diagnostic interface [17] | 394 preterm (<35 w PMA) | 8 | 1 | 1 (Line length) | U; Clu | n/a |
| Burst detection in neonatal EEG [18] | 10 preterm + 10 full-term | 1 | n/a | 1 (Envelope derivative operator) | S; Th | AUC≥0.9 |
| EEG differentiation after asphyxia [19] | 34 full-term | 12 | 1 | 3 (Amplitude and time based) | S; SVM | Acc=84% |
| Automated detection of bursts and IBIs [20] | 36 preterm (<30 w GA) | 8 | 3 | 1 (Nonlinear energy operator) | S; Th | Algorithm/Rater: Acc=81%, $\kappa$=0.63 ; Inter-rater: Acc=71% $\kappa$=0.58 |
| Burst/IBI classification by age [21] | 26 extremely preterm | 2 | 0 | 1 (Range EEG) | U; Th | n/a |

Table 1: Summarized review of burst detection methods in preterm infants and other populations. Abbreviations not defined in the body text are: #Chan, number of channels; #Sc, number of scorers; w, weeks; PMA, post-menstrual age; GA, gestational age; S, supervised; U, unsupervised; FLD, Fisher linear discriminant; SVM, support vector machines; Clu, clustering; n/a not available; Acc, accuracy; Sn, sensitivity; AUC, area under receiver operating characteristic (ROC) curve; $\kappa$, Cohen's kappa.

## 2.3 Gold Standard

The gold standard, i.e. reference labels to train and test the classification algorithms, was generated by merging the visual assessments from the annotated dataset. This procedure involved two steps: 1) the computation of intra-rater marks from two repetitions and 2) the computation of the inter-rater marks.

### 2.3.1 Intra-rater marks

The two neonatologists (raters A and B) were trained to use a computer program designed purposely to mark the bursts limits in a screen showing 20-second windows of pre-processed EEG (15 windows for each of the 14 infants). This procedure was repeated twice by each rater, in different days to avoid bias.

The annotations were then converted to discrete, binary series that coded bursts with ones and IBIs with zeros. Each category was associated, respectively with $Class\ 1$ and $Class\ 0$. We thus obtained four binary arrays $Y_{r,i} \in \{0,1\}$ of length $L = D \times F_s$, where $r = \{A, B\}$ represents the rater's code and $i = \{1, 2\}$ is the repetition number.

Intra-rater marks were finally generated by merging the marks for each rater, $Y_r$, which included the bursts of the two replicates [12]:

$$Y_r = y_r(k) = \begin{cases} 0 & \text{if } y_{r,1}(k) + y_{r,2}(k) = 0 \\ 1 & \text{otherwise} \end{cases} \quad (1)$$

for $k = 1, ..., L$ and $r = \{A, B\}$.

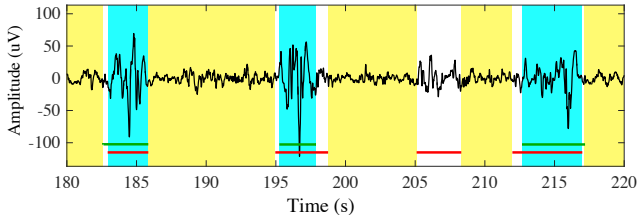


Figure 1: Example of EEG scored with the raters labels. Green and red lines are $Y_A$, $Y_B$ respectively. Only the consensual marks are taken into account to establish the gold standard (blue areas, bursts; yellow areas, IBIs). Areas in white represent disagreeing zones.

#### 2.3.2 Inter-rater marks

We defined the gold standard, $Y$, as in [12], i.e. the unanimous decisions between intra-raters marks. Intervals without agreement were not considered and labeled as empty values ($\emptyset$):

$$Y = y(k) = \begin{cases} 1 & \text{if } \prod_r y_r(k) = 1 \\ 0 & \text{if } \sum_r y_r(k) = 0 \\ \emptyset & \text{otherwise} \end{cases} \quad (2)$$

with $k = 1, \ldots, L$ and $r = \{A, B\}$. An example is given in Fig. 1.

# 3 Methods

## 3.1 Burst detection framework

To test the different burst detectors, we employed the general scheme depicted in Fig. 2 that can be divided in three main blocks: pre-processing, feature extraction and classification. The content of some blocks depends on the classification approach.

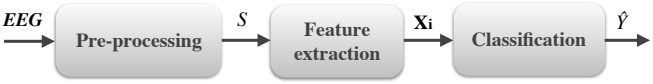


Figure 2: Block diagram of the employed framework to detect bursts. The input signals ($EEG$) come from a two channel system. The pre-processing block yields a one dimensional signal, $S$, from which a feature vector, $\mathbf{X}_i$, is extracted in each window $i$. The classification block outputs a binary, one-dimensional signal, $\widehat{Y}$, with the predicted bursts.

#### Pre-processing

In this block, signal-to-noise ratios in EEG signals are improved by applying artifact correction or rejection and filters. Due to prone position and nursing, certain artifacts are typically more abundant in one channel. Here, the less contaminated one is selected regarding its statistical properties (variance, kurtosis, joint probability of EEG activity values [29]). Then, a linear phase band-pass filter whose respective lower and upper cut-off frequencies are set to $F_l$ and $F_u$, is applied. Cut-off frequencies can vary depending if baseline and high frequency noise needs to be attenuated or if a specific bandwidth wants to be enhanced. The resulting signal, $S$, is finally sub-sampled to 128 Hz.

#### Feature extraction

This block computes a number of functions on $S$ to obtain a feature vector $\mathbf{X_i} \in \mathbb{R}^{N_f}$ with $N_f$ the number of features, and $i = 1, \ldots N_w$, with $N_w$ the number of windows of $S$. Features, chosen to capture pronounced characteristics in bursts (see Table 2), have already been exploited to solve EEG classification problems. Since $\mathbf{X_i}$ are computed in overlapping $W$-second windows, the effective sampling rate of features with respect to $S$ is reduced. Hence slow trends were enhanced over fast transients by performing a smoothing of $\mathbf{X_i}$ as suggested by [11]. To this purpose we applied a 1-dimensional, 10th order median filter on each feature [30]. Finally, to avoid outliers that might decrease classification performances, $\mathbf{X_i}$ was quantile normalized to impose their values to fall within the 1st and 99st percentiles [11].

#### Classification

The purpose of this block is to identify the labels from new observations using $\mathbf{X_i}$. The proposed classifier based on logistic regression and its competitors are described below. Provided that short bursts or IBIs rarely exist, the output of the classifiers were also smoothed to improve the performance of detections. Hence, the output of this block, $\widehat{Y}$, was finally obtained by removing isolated events below a given time in seconds, $t_B$, applying a filtering procedure similar than [11].

## 3.2 Classification based on logistic regression

Predictive models based on logistic regression has been successfully employed in a variety of biomedical domains [32–34]. Unlike binary classifiers, that are purely dichotomous, LR provides the class probability for one of the two categories.

In logistic regression [35], the class probability $\pi_i$ is expressed through a function called logit, related to the feature vector $\mathbf{X_i}$ :

$$logit(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = w_0 + \boldsymbol{w} \cdot \mathbf{X_i}. \quad (3)$$

| Name | Description |
|---|---|
| $Mm$ | Difference between the maximum and the minimum value |
| $DM$ | Maximum of absolute values of the discrete difference: $DM = \max_{k=1,..,l}\{\|S(k) - S(k-1)\|\}$, where $l$ is the number of points in $W$ |
| $SD$ | Standard deviation |
| $Kt$ | Kurtosis |
| $NL$ | Nonlinear energy operator (NLEO) [9]: $NL = \frac{1}{l}\sum_{k=1}^{l} S(k)S(k-3) - S(k-1)S(k-2).$ |
| $AD$ | Averaged differentiation, defined as: $AD = \frac{1}{l}\sum_{k=1}^{l} \|S(k) - S(k-1)\|.$ |
| $Hs$ | Shannon Entropy [11]: $Hs = -\sum_{q} p(I_q)\log p(I_q),$ where $p(I_q)$, $q = 1...Q$ is a discrete set of probabilities estimated by counting the $l$ points within $Q$ =16 histogram bins. |
| $Pw$ | Power between 0.5 – 3Hz, estimated by an autoregressive model using the Burg method. Model order (15) was set to the mean value provided by Akaike's information criterion [31]. |

Table 2: Definition of the features applied on each EEG window ($W$ seconds) for multi-feature classifiers.

where $\boldsymbol{w} = [w_1, .., w_d]$ is the vector of regression coefficients and $w_0$ is the intercept. The inverse of the above expression, called logistic function, is expressed as:

$$logit^{-1}(\pi_i) = \frac{1}{1 + e^{-(w_0 + \boldsymbol{w}\mathbf{X_i})}} = g(\mathbf{X_i}, \boldsymbol{w}). \quad (4)$$

An important characteristic of the logistic function is that it is bounded between 0 and 1, and thus, it can be used directly to estimate the probabilities of the possible outcomes as $P(Y = 1|\boldsymbol{w}, \mathbf{X_i}) = g(\mathbf{X_i}, \boldsymbol{w})$.

Given the pair of features and labels $\{\mathbf{X_i}, Y_i\}$, the learning process aims at finding the best $\boldsymbol{w}$, which is to maximize the conditional probabilities $P(Y_i|\mathbf{X_i}, \boldsymbol{w})$ [35]. This can be achieved by the maximum likelihood estimation (MLE) method. We employed the Newton-Raphson's hill-climbing algorithm, an iterative procedure that maximizes the log likelihood function until a convergence criterion (coefficients leading to the most accurate predictions) is reached.

Once the optimal coefficients, $\widehat{\boldsymbol{w}}$, are obtained, class probabilities, $\widehat{\pi}_i$, are provided by the logistic function. The class membership is decided by a cut-off value $c$ such that $f(\widehat{\pi}_i) > c$ assigns the predictive output value, $\widehat{y}$, to Class 1, and $f(\widehat{\pi}_i) \leq c$ assigns $\widehat{y}$ to Class 0. Here, we fixed $c$ to 0.5.

## 3.3 Alternate multi-feature classifiers

In this paper, we have also evaluated the detection of bursts using three widely employed multi-feature, supervised classifiers suitable for binary classification problems: Linear discriminant analysis, support vector machines and the K-nearest neighbor technique. They are briefly described below.

### 3.3.1 Linear discriminant analysis

Linear discriminant analysis can be applied to solve two-class classification problems simply and efficiently based on the characteristics of each class (mean, covariance matrix) [36]. The LDA classifier finds a discriminant function, i.e. the linear combination of the multi-dimensional features that best separates the two classes. This function provides scores for each class, being the highest values associated to more likely classes.

### 3.3.2 Support vector machines

The SVM is a very popular machine learning technique used in a variety of applications [37]. This classifier uses a transformation (kernel) function to project the data into a higher dimensional space, where classes may become linearly separable. More versatile than linear kernel functions, we used a Gaussian radial basis function (RBF) to guarantee the existence of a non-linear decision boundary:

$$K(x_i, x_j) = \exp^{\|x_i - x_j\|^2/\sigma}, \quad (5)$$

where $x_i$ and $x_j$ denote two feature vectors and the kernel parameter $\sigma$ is the radius of influence of the learning samples selected as support vectors by the model. The other parameter in SVMs, the weight of the soft margin cost function ($C$) [38], needs to be adjusted for an optimal decision boundary. While small values provides "local" solutions over-fitting the model, high values tend to simplify boundaries and may not provide accurate separations. Both parameters $\sigma$ and $C$ were optimized by the sequential minimal optimization method (SMO) [39].

### 3.3.3 K-nearest neighbor

The kNN is a nonparametric and nonlinear classifier based on proximity criteria. Given the training set of features, the algorithm identifies the $k$ closest neighbor vectors to classify a new instance. The class assigned to

the new instance is then decided by majority vote, i.e. the class accounting for more neighbors. The value of $k$ was set as the square root of the number of instances [40].

### 3.4 Detection by thresholding

Thresholding is a simple technique that can be employed when one-dimensional feature vectors $\mathbf{X}$ can be partitioned in two disjoint regions (classes) by a threshold $T$. To find $T$, an optimization procedure that maximizes the agreements with the gold standard is performed. New instances are then classified by a simple rule: if the feature value exceeds $T$, it is labeled as $Class$ 1, otherwise as $Class$ 0.

We employed the thresholding approach proposed by Palmu et al. [12]. Briefly, it consists on first preprocessing EEG by a band-pass filter with cut-off frequencies $F_l$ and $F_u$, respectively. Next, the feature (given by the NLEO operator) is computed in $W$-second windows so that values over $T$ provided a first classification, corrected in a second instance by eliminating bursts below $t_B$. By means of an iterative process, $F_l$, $F_u$, $W$, $T$ and $t_B$ were optimized to obtain a maximum agreement with their gold standard. We simplified this procedure by optimizing $T$ and imposing the remaining parameters (see Section 4.1).

### 3.5 Measures of agreement and performance

To assess the degree of agreement within human raters and between human and automatic classifications, we employed the Cohen's kappa coefficient [41]:

$$\kappa = \frac{P_o - P_c}{1 - P_c}, \tag{6}$$

where $P_o$ is the observed agreement among raters (the proportion of windows where the observers agreed) and $P_c$ is the probability expected by chance. The upper limit of this statistic ($\kappa = 1$) occurs only when there is perfect agreement. The lower limit ($\kappa \leq 0$) depends on the marginal distributions and occurs when agreements are due to chance [41].

The performance of the classifiers was evaluated by accuracy and receiver operating characteristic (ROC) curves. Accuracy ($Acc$) is defined as the percentage of windows correctly classified over the total number of windows in each labeled EEG. ROC curves represent a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the ROC curves (AUC) summarizes the overall ability of the classifiers to discriminate between the two classes and ranges from 0.5 (random classification) to 1 (perfect classification).

To obtain unbiased estimations of accuracy and AUC, the performance of the classifiers was examined in infants that did not take part in the training process. Hence, leave-one-out cross-validation (LOOCV) was applied:

1. Form a *validation* subset by selecting one segment from the annotated dataset.

2. Build the classification model with the *training* subset, i.e. the remaining $N_e$-1 segments .

3. Test the *validation* subset with the trained model.

4. Repeat the above steps until each of the $N_e$ segments has been omitted and tested once.

## 4 Results

### 4.1 Setting up automatic detections

Filter cut-off frequencies $F_l$ and $F_u$ were set to 0.1 – 30 Hz for multi-feature classifiers. For thresholding, these values were modified (0.5 – 8 Hz) to meet de requierements of [24]. In all cases, features were computed by 75% overlapping windows of $W$=1 second. This choice is justified by the minimal duration of the bursts in the gold standard but also by a trade-off between reasonable resolution (0.25 s) and computational time. The minimal burst time, $t_B$ was set to 1 second.

We proceeded then to select the most relevant features for LDA, SVM, kNN and LR. Feature matrix was composed, per each infant, by 1197 rows (data points) and 8 columns (features). Given that the number of features is low with respect to the number of observations, a wrapper feature selection method was employed. The most relevant features were retained by sequential forward selection (SFS), i.e. subsets of features are iteratively combined based on the classifier performance until a maximum is reached. The maximal performance was evaluated by the mean accuracy yielded by LOOCV. The number of retained features obtained by SFS depended on the classification method. While LDA reached the best accuracy using only two features ($Mm$, $Kt$), SVM needed all excepting $Pw$. For the kNN method, five features were selected ($Mm$, $SD$, $NL$, $AD$, $Hs$) and LR retained three features ($Mm$, $SD$, $NL$). Of note, $Pw$ were discarded by all classifiers, suggesting that it may be redundant or poorly correlated with the labels. On the other hand, $Mm$ constituted the most relevant feature as it was selected in all cases.

The labels for the classifiers were provided by the gold standard which, in summary, consisted on 70 minutes of labeled EEG with 311 bursts and 318 IBIs.

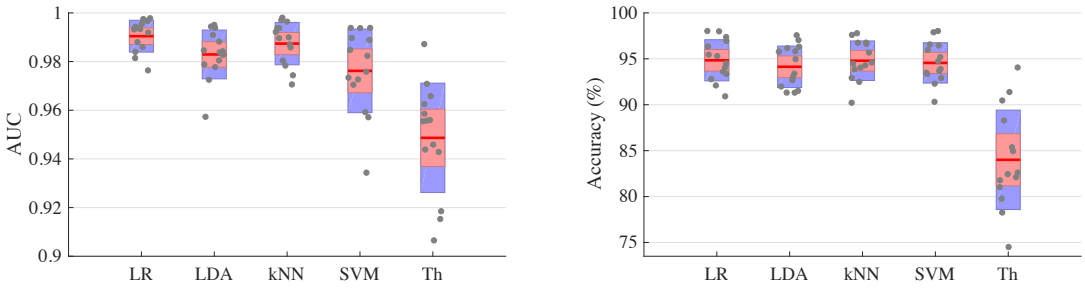


Figure 3: Performance of all classifiers in terms of areas under ROC curves (AUC, left panel) and accuracy after performing feature selection. Horizontal red lines denote mean values, red and blue zones represent standard deviations and 95% confidence intervals (C.I), respectively. Individual values are given by grey circles. Only the performance of $Th$ is significantly below the rest of classifiers (non-overlapping C.I.).

## 4.2 Comparison of automatic detections

The performance of the classifiers in terms of AUCs and accuracies are depicted in Fig. 3. Concerning multi-feature classifiers, accuracies were almost identical by using LR, SVM and kNN ($Acc \approx 95\%$). Little differences exist between these methods when comparing confidence intervals and dispersion. The LDA was slightly below (94%). LR resulted computationally simpler (uses only 3 features), faster and more intuitive method than the other classifiers as it provides directly the probability of burst (ease of setting a working point by simply changing the cut-off value $c$).

Our results revealed that thresholding performed poorly ($Acc$=84%) compared to other works detecting bursts on more immature infants (23 - 28 w PMA), with $Acc$=90% in average [12]. Indeed, the single feature employed by this algorithm does not describe properly the EEG complexity in older preterm infants. Hence, the use of a multi-feature classifier that includes additional EEG descriptors is necessary to improve burst detection in older populations.

Even if accuracies provided by LR, SVM and kNN are in the same levels of some of the existing burst supression detectors in full-terms [13] and above burst/IBI classifiers for preterms [14, 16], performances should be compared with caution as they are subject to the design of the gold standard. Thus, the comparison of automatic detections with those obtained by human raters will provide a more realistic idea of the behavior of the classifiers.

## 4.3 Visual *vs.* automatic detections

The annotated dataset also served to compare the agreements within raters and between raters and the classifiers. For human observations, kappa coefficients were, in average, equal to 0.62. This result improves reported values in populations <30 weeks PMA [20] (mean $\kappa$=0.58).

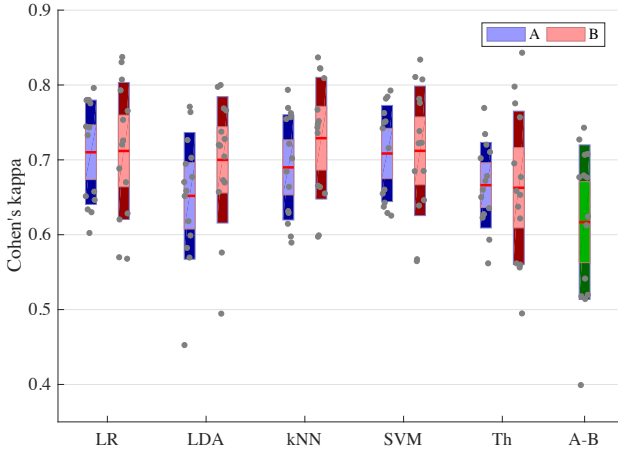


Figure 4: Comparison of human observations against the automatic detections provided by the five tested classifiers. Blue and red boxes represent kappa coefficients with raters A and B, respectively. None of the A-B pairs showed statistically significant differences in a Mann-Whitney U test. Right green box shows inter-rater agreements. Boxes read as in Fig. 3.

In terms of accuracy, our mean agreement equals 81%, a satisfactory result if compared to values obtained in younger cohorts, for instance 81% in 28 to 30 w PMA infants [12] or 80% in 29 to 34 w PMA [14]. In our experiment, discordance was mainly found at the beginning and end of bursts and in few cases concerned a entire burst.

Comparing the kappa coefficients in Fig. 4, it can be stated that automatic-human values are increased with respect to human-human rates. This can be explained by the fact that the gold standard used to train the classifiers is an intermediate reference, i.e. from raters unanimous decisions. Both LR and SVM yielded best averaged human-automatic agreements ($\kappa$=0.71, $Acc$=86%), but for computational efficiency, LR was our method of choice for the study of maturation presented in Section

4.4.

### 4.3.1 Discontinuity parameters

In neonatology, maturational patterns are often assessed from the quantitative analysis of EEG bursts. Here, we compared the following measures, also referred to as discontinuity parameters:

- Number of bursts per minute ($N_{Bm}$)
- Mean duration of bursts ($\bar{t}_B$)
- Mean duration of IBIs ($\bar{t}_I$)
- Maximal duration of IBIs ($t_{I,max}$)

As it can be observed in Fig. 5, values from automated detections are intermediate to those obtained by the raters, excepting $N_{Bm}$ (whose median is over the values obtained by manual marks). Regarding this parameter, differences between LR and B were statistically significant whereas differences between LR and A were not. Significant differences concerning the rest of comparisons with LR cannot be considered relevant as there were also significant differences between A and B (see horizontal lines in Fig. 5). Therefore, automatic detections can be, in general, comparable to human judgment.

## 4.4 Study of maturation in a non-annotated test dataset

To have a qualitative idea of the infant's maturation in our database, we computed the above described discontinuity parameters using the most performing burst classification model. Infants having sufficiently long periods in quiet sleep (>300 seconds) were selected from the main database of 31 infants, discarding too short, unstable sleep patterns. Hence, $N_t$=21 EEGs from 21 different infants, summing up approximately two hours of EEG signals, constituted a non-annotated test dataset.

Then, we divided the test dataset in two subsets of groups according to the median PMA and weight. This allowed to compare the degree of maturity by age (group $G_1^{PMA}$ = [36, 38.2] versus group $G_2^{PMA}$ = [38.2, 40] w. PMA) and by weight (group $G_1^W$ = [1.36, 2.50] versus $G_2^W$ = [2.50, 2.86] Kg). The four discontinuity parameters before described plus the percentage of bursts ($\%_B$) were calculated from the detections yielded by the LR classifier (see Table 3).

Significant differences were found in certain parameters regarding weight or age groups. In general, patterns tend to be more continuous as evidenced by the increase of $\bar{t}_B$ and $\%_B$ or the reduction of $t_{I,max}$ in more mature groups. These changes are in concordance with widely accepted maturational criteria, such as the IBI reduction and the prolongation of bursts with increasing PMA

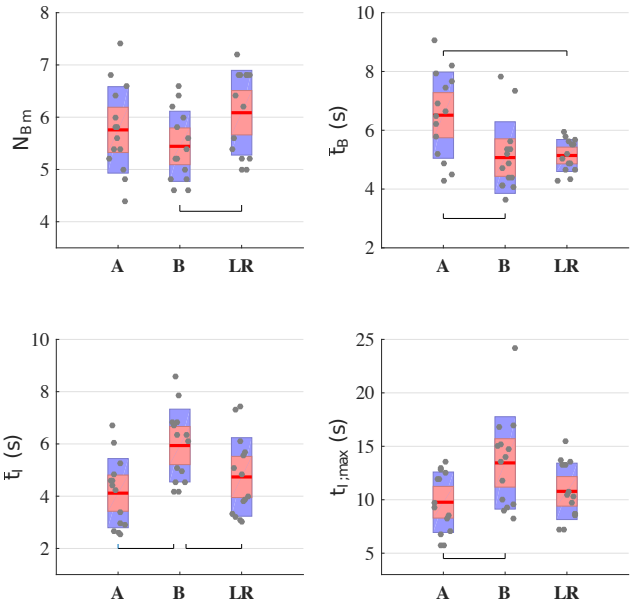


Figure 5: Characteristics of bursts (discontinuity parameters) according to the raters (A, B) and automatic detection by logistic regression (LR). From left to right and top to down: $N_{Bm}$ (number of bursts per minute); $\bar{t}_B$ (mean duration of bursts), $\bar{t}_I$ (mean duration of IBIs) and $t_{I,max}$ (maximal duration of IBIs). Horizontal lines grouping a pair of boxes denote statistically significant differences ($p$ <0.05) in a Mann-Whitney U test. Interpretation of boxes as in Fig. 3.

| | PMA (weeks) | | Weight (Kg) | |
|---|---|---|---|---|
| | 36-38.2 | 38.2-40 | 1.36-2.50 | 2.50-2.86 |
| $N_{Bm}$ | 5.6 ± 0.3 | 5.9 ± 1.1 | 6.0 ± 0.9 | 5.4 ± 0.5 |
| $\bar{t}_B$ (s) | 5.5 ± 0.5 | 6.0 ± 1.5 | 5.2 ± 0.7 | 6.4 ± 1.2* |
| $\bar{t}_I$ (s) | 5.3 ± 0.8 | 4.3 ± 0.6* | 4.7 ± 0.8 | 4.6 ± 0.9 |
| $t_{I,max}$ (s) | 14 ± 0.8 | 11 ± 4.1 | 14 ± 1.8 | 11 ± 3.4* |
| $\%_B$ | 52 ± 5.9 | 57 ± 6.5 | 52 ± 4.3 | 57 ± 8.1* |

Table 3: Discontinuity parameters versus PMA and weight in the non-annotated test dataset of 21 infants. Parameters are: $N_{Bm}$ (number of bursts per minute); $\bar{t}_B$ (mean duration of bursts), $\bar{t}_I$ (mean duration of IBIs) and $t_{I,max}$ (maximal duration of IBIs). Asterisks denote statistically significant differences ($p < 0.05$) in a Mann-Whitney U test.

[1,23]. Moreover, infants between 35 and 39 weeks PMA rarely exhibit IBIs exceeding 20 seconds, and their mean durations range 4 to 10 seconds depending on the sleep state [2,42], two descriptions corroborated by our results.

# 5 Discussion

The results of the comparative study performed on the annotated dataset of 14 infants showed that the proposed multi-feature classifiers improved the widely em-

ployed thresholding approach and most of the available multi-feature approaches in the literature. Indeed, our results benefit from the selection of the channel having the lowest degree of artifacts and from the inclusion of the appropriate burst descriptors by the most relevant features.

The logistic regression method reached the highest accuracy rate (95%) and mean AUC (0.99). Not far, the k-nearest neighbor technique and support vector machines were alternatives with similar performances. Nevertheless, these performances do not take into account the possible classification errors in disagreeing zones since the gold standard used to train the algorithms was build from consensual annotations [12]. As a result, accuracies and AUCs might be over-estimated.

Indeed, this uncertainty is implicit when consensus-based gold standards are adopted. On the one hand, the inclusion of two raters approaches the gold standard to the "truth". On the other hand, disagreement zones appear, reducing the reliability of evaluations. Hence the importance of a complementary evaluation comparing automatic and human detections. In this comparison, the LR-based classifier and SVM yielded the best human-automatic agreements, with an average accuracy of 86%.

Effectively, results from LR classification indicate that parameters describing the discontinuity of bursts are in the same range than clinicians' judgments. Therefore, the implementation of this detector would help assessing the infant's maturity in a more repeatable, faster and cost-effective way.

Concerning the computational cost of the algorithms, we also found LR advantageous with respect the other multi-feature classifiers during the testing process. Despite the differences were considerable (LR was 1000 times faster than kNN and 10 faster than LDA), the slowest classifier needed less than 0.1 seconds to classify 300 seconds of EEG signals (using a 2,8 GHz Intel Core i7 processor with 16 megabytes of RAM memory). Therefore, from the computational point of view, the use of any of the multi-feature solutions should not be critical in a real-time implementation of the burst detector.

In an attempt to study the maturation in a larger cohort, we applied the LR classifier in a test database composed by 21 recordings. The discontinuity parameters obtained by our classifier showed the normal evolution of electroencephalographic patterns [1, 23] comparing age and weight grouped infants. Although indicative, the study of maturation using the trained model on 14 records should be interpreted with caution as discontinuity parameters are inferred from different infants at specific maturity levels. Including more infants to the annotated dataset could be a possible solution to improve the predictive power of the classifiers. However, only a horizontal database would ensure a reliable assessment of intra-individual maturity. In future studies, new recordings will be included to improve the proposed burst detector.

# 6 Conclusion

This study shows that the EEG burst detection problem in very preterm infants who reached term age can be successfully addressed using multi-feature classification on a single EEG channel. The main advantage of our proposal relies on its simplicity, reliability and computational efficiency thanks to a logistic regression detector. This framework could add new functionality to current bedside monitors, but also it could open the way to home monitors (integrating wearable devices or EEG portable headsets) to follow up maturation in preterm infants after hospital discharge.

# Acknowledgments

Authors thank the clinicians from the CHU of Rennes for their involvement in this study. The research for this paper was financially supported by the project INTEM between Rennes University Hospital and the LTSI - INSERM U1099.

# References

[1] J. S. Hahn, H. Monyer, and B. R. Tharp, "Interburst interval measurements in the EEGs of premature infants with normal neurological outcome," *Electroencephalography and Clinical Neurophysiology*, vol. 73, no. 5, pp. 410–418, 1989.

[2] M. Hayakawa, A. Okumura, F. Hayakawa, K. Watanabe, M. Ohshiro, Y. Kato, R. Takahashi, and N. Tauchi, "Background electroencephalographic (EEG) activities of very preterm infants born at less than 27 weeks gestation: a study on the degree of continuity," *Archives of Disease in Childhood. Fetal and Neonatal Edition*, vol. 84, pp. F163–F167, May 2001.

[3] M. Scher, D. Steppe, R. Dahl, S. Asthana, and R. Guthrie, "Comparison of EEG sleep measures in healthy full-term and preterm infants at matched conceptional ages," *Sleep*, vol. 15, no. 5, pp. 442–448, 1992.

[4] M. André, M.-D. Lamblin, A.-M. d'Allest, L. Curzi-Dascalova, F. Moussalli-Salefranque, S. N. Tich, M.-F. Vecchierini-Blineau, F. Wallois, E. Walls-Esquivel, and P. Plouin, "Electroencephalography in premature and full-term infants. Developmental features and glossary," *Clinical Neurophysiology*, vol. 40, no. 2, pp. 59–124, 2010.

[5] M. Lamblin, M. Andre, M. Challamel, L. Curzi-Dascalova, A. d'Allest, E. De Giovanni, F. Moussalli-Salefranque, Y. Navelet, P. Plouin, M. Radvanyi-Bouvet, *et al.*, "Electroencephalography of the premature and term newborn. Maturational aspects and glossary," *Clinical neurophysiology*, vol. 29, no. 2, pp. 123–219, 1999.

[6] T. Randò, D. Ricci, R. Luciano, M. F. Frisone, G. Baranello, T. Tonelli, M. Pane, C. Romagnoli, G. Tortorolo, E. Mercuri, *et al.*, "Prognostic value of EEG performed at term age in preterm infants," *Child's Nervous System*, vol. 22, no. 3, pp. 263–269, 2006.

[7] N. Hayashi-Kurahashi, H. Kidokoro, T. Kubota, K. Maruyama, Y. Kato, T. Kato, J. Natsume, F. Hayakawa, K. Watanabe, and A. Okumura, "EEG for predicting early neurodevelopment in preterm infants: an observational cohort study," *Pediatrics*, vol. 130, no. 4, pp. e891–e897, 2012.

[8] J. Turnbull, K. Loparo, M. Johnson, and M. Scher, "Automated detection of tracé alternant during sleep in healthy full-term neonates using discrete wavelet transform," *Clinical Neurophysiology*, vol. 112, no. 10, pp. 1893–1900, 2001.

[9] M. Särkelä, S. Mustola, T. Seppänen, M. Koskinen, P. Lepola, K. Suominen, T. Juvonen, H. Tolvanen-Laakso, and V. Jäntti, "Automatic analysis and monitoring of burst suppression in anesthesia," *Journal of Clinical Monitoring and Computing*, vol. 17, pp. 125–134, Feb. 2002.

[10] K. Pfurtscheller, G. R. Müller-Putz, B. Urlesberger, W. Müller, and G. Pfurtscheller, "Relationship between slow-wave EEG bursts and heart rate changes in preterm infants," *Neuroscience Letters*, vol. 385, pp. 126–130, Sept. 2005.

[11] J. Löfhede, N. Löfgren, M. Thordstein, A. Flisberg, I. Kjellmer, and K. Lindecrantz, "Classification of burst and suppression in the neonatal electroencephalogram," *Journal of Neural Engineering*, vol. 5, p. 402, Dec. 2008.

[12] K. Palmu, S. Wikström, E. Hippeläinen, G. Boylan, L. Hellström-Westas, and S. Vanhatalo, "Detection of 'EEG bursts' in the early preterm EEG: visual vs. automated detection," *Clinical Neurophysiology*, vol. 121, pp. 1015–1022, July 2010.

[13] J. Löfhede, M. Thordstein, N. Löfgren, A. Flisberg, M. Rosa-Zurera, I. Kjellmer, and K. Lindecrantz, "Automatic classification of background EEG activity in healthy and sick neonates," *Journal of neural engineering*, vol. 7, no. 1, p. 016007, 2010.

[14] W. Jennekens, L. S. Ruijs, C. M. Lommen, H. J. Niemarkt, J. W. Pasman, V. H. van Kranen-Mastenbroek, P. F. Wijn, C. van Pul, and P. Andriessen, "Automatic burst detection for the EEG of the preterm infant," *Physiological measurement*, vol. 32, no. 10, p. 1623, 2011.

[15] V. Matic, P. J. Cherian, K. Jansen, N. Koolen, G. Naulaers, R. M. Swarte, P. Govaert, G. H. Visser, S. Van Huffel, and M. De Vos, "Automated EEG interburst interval detection in neonates with mild to moderate postasphyxial encephalopathy," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 17–20, 2012.

[16] N. Koolen, K. Jansen, J. Vervisch, V. Matic, M. De Vos, G. Naulaers, and S. Van Huffel, "Automatic burst detection based on line length in the premature EEG," in *BIOSIGNALS*, pp. 105–111, 2013.

[17] P. E. Chauvet, S. N. T. Tich, D. Schang, and A. Clément, "Evaluation of automatic feature detection algorithms in EEG: Application to interburst intervals," *Computers in biology and medicine*, vol. 54, pp. 61–71, 2014.

[18] J. M. O'Toole, A. Temko, and N. Stevenson, "Assessing instantaneous energy in the EEG: a non-negative, frequency-weighted energy operator," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3288–3291, 2014.

[19] V. Matic, P. J. Cherian, N. Koolen, G. Naulaers, R. M. Swarte, P. Govaert, S. Van Huffel, and M. De Vos, "Holistic approach for automated background EEG assessment in asphyxiated full-term infants," *Journal of neural engineering*, vol. 11, no. 6, p. 066007, 2014.

[20] K. Murphy, N. J. Stevenson, R. M. Goulding, R. O. Lloyd, I. Korotchikova, V. Livingstone, and G. B. Boylan, "Automated analysis of multi-channel EEG in preterm infants," *Clinical Neurophysiology*, vol. 126, no. 9, pp. 1692–1702, 2015.

[21] M. A. Navakatikyan, D. O'Reilly, and L. J. Van Marter, "Automatic measurement of interburst interval in premature neonates using range EEG," *Clinical Neurophysiology*, vol. 127, no. 2, pp. 1233–1246, 2016.

[22] R. Brenner, "The electroencephalogram in altered states of consciousness," *Neurologic clinics*, vol. 3, no. 3, pp. 615–631, 1985.

[23] E. Biagioni, L. Bartalena, A. Boldrini, G. Cioni, S. Giancola, and A. E. Ipata, "Background EEG activity in preterm infants: correlation of outcome with selected maturational features," *Electroencephalography and clinical neurophysiology*, vol. 91, pp. 154–162, Sept. 1994.

[24] K. Palmu, N. Stevenson, S. Wikström, L. Hellström-Westas, S. Vanhatalo, and J. M. Palva, "Optimization of an NLEO-based algorithm for automated detection of spontaneous activity transients in early preterm EEG," *Physiological Measurement*, vol. 31, pp. N85–93, Nov. 2010.

[25] J. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *International Conference on Acoustics, Speech, and Signal Processing, 1990 (ICASSP-90)*, pp. 381–384 vol.1, 1990.

[26] R. Agarwal and J. Gotman, "Adaptive segmentation of electroencephalographic data using a nonlinear energy operator," in *Proceedings of the 1999 IEEE International Symposium on Circuits and Systems (ISCAS'99)*, vol. 4, pp. 199–202, 1999.

[27] T. Mialet-Marty, A. Beuchée, W. Ben Jmaa, N. N'guyen, X. Navarro, F. Porée, A. M. Nuyt, and P. Pladys, "Possible predictors of cardiorespiratory events after immunization in preterm neonates," *Neonatology*, vol. 104, no. 2, pp. 151–155, 2013.

[28] H. F. Prechtl, "The behavioural states of the newborn infant (a review)," *Brain research*, vol. 76, no. 2, pp. 185–212, 1974.

[29] A. Delorme, T. Sejnowski, and S. Makeig, "Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis," *Neuroimage*, vol. 34, no. 4, pp. 1443–1449, 2007.

[30] G. R. Arce, *Nonlinear signal processing: a statistical approach*. John Wiley & Sons, 2005.

[31] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*, pp. 199–213, 1998.

[32] M. Hajmeer and I. Basheer, "Comparison of logistic regression and neural network-based classifiers for bacterial growth," *Food Microbiology*, vol. 20, no. 1, pp. 43–55, 2003.

[33] J. Liao and K.-V. Chin, "Logistic regression for disease classification using microarray data: model selection in a large p and small n case," *Bioinformatics*, vol. 23, no. 15, pp. 1945–1951, 2007.

[34] B. D. Ripley, *Pattern recognition and neural networks*. Cambridge university press, 2007.

[35] D. W. Hosmer Jr and S. Lemeshow, *Applied logistic regression*. John Wiley & Sons, 2004.

[36] G. McLachlan, *Discriminant analysis and statistical pattern recognition*, vol. 544. John Wiley & Sons, 2004.

[37] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[38] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[39] R.-E. Fan, P.-H. Chen, and C.-J. Lin, "Working set selection using second order information for training support vector machines," *Journal of machine learning research*, vol. 6, no. Dec, pp. 1889–1918, 2005.

[40] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[41] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[42] M.-F. Vecchierini, M. André, and A. d'Allest, "Normal EEG of premature infants born between 24 and 30 weeks gestational age: Terminology, definitions and maturation aspects," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 37, no. 5, pp. 311–323, 2007.