KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Ansari, A.H., Cherian, P.J., Dereymaeker, A., Matic, V., Jansen, K., De Wispelaere, L., Dielman, C., Vervisch, J., Swarte, R.M., Govaert, P., Naulaers, G., De Vos, M., Van Huffel, S., **Improved multi-stage neonatal seizure detection using a heuristic classifier and a data-driven post-processor** Clin. Neurophysiol, vol. 127, 2016, 3014-3016.
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	http://www.sciencedirect.com/science/article/pii/S1388245716304485
Journal homepage	http://www.journals.elsevier.com/clinical-neurophysiology
Author contact	your email amirhossein.ansari@kuleuven.be your phone number *+ 32 (0)16 326148*
IR	url in Lirias https://lirias.kuleuven.be/handle/123456789/544129

(article begins on next page)

Improved multi-stage neonatal seizure detection using a heuristic classifier and a data-driven post-processor

A. H. Ansari, P. J. Cherian, A. Dereymaeker, V. Matic, K. Jansen, L. De Wispelaere, C. Dielman, J. Vervisch, R.M. Swarte, P. Govaert, G. Naulaers, M. De Vos, S. Van Huffel

Highlights

- An improved neonatal seizure detection method is discussed.
- A set of characteristic features of seizures are identified by data-driven methods.
- Described core characteristics of neonatal seizures can easily be used for other automated methods.

Keywords:

Automated neonatal seizure detection; Hypoxic-ischemic encephalopathy; Machine learning; Support vector machines

Abstract

Objective: After identifying the most seizure-relevant characteristics by a previously developed heuristic classifier, a data-driven post-processor using a novel set of features is applied to improve the performance.

Methods: The main characteristics of the outputs of the heuristic algorithm are extracted by five sets of features including synchronization, evolution, retention, segment, and signal features. Then, a support vector machine and a decision making layer remove the falsely detected segments.

Results: Four datasets including 71 neonates (1023 hours, 3493 seizures) recorded in two different university hospitals, are used to train and test the algorithm without removing the dubious seizures. The heuristic method resulted in a false alarm rate of 3.81 per hour and good detection rate of 88% on the entire test databases. The post-processor, effectively reduces the false alarm rate by 34% while the good detection rate decreases by 2%.

Conclusion: This post-processing technique improves the performance of the heuristic algorithm. The structure of this post-processor is generic, improves our understanding of the core visually determined EEG features of neonatal seizures and is applicable for other neonatal seizure detectors.

Significance: The post-processor significantly decreases the false alarm rate at the expense of a small reduction of the good detection rate.

A. Introduction

Seizures are a common and distinctive sign of serious brain dysfunction in neonates (Volpe 2008). The majority of neonatal seizures have an acute symptomatic basis and one of the most important causes is hypoxic ischemic encephalopathy (HIE) (Hahn and Olson 2004; Cherian et al. 2011). Clinical presentation can be highly variable and manifestations of neonatal seizures can be subtle, absent or resemble normal behavior. It is known that after treating with anticonvulsants, clinical seizures will change in subclinical seizures (Connell et al. 1989; Scher et al. 2003). Hence, clinical observation alone is ill-suited for their identification and monitoring (Bye and Flanagan 1995; Rennie et al. 2004; Murray et al. 2008). Monitoring of the electroencephalogram (EEG) along with video is the gold standard for diagnosing and monitoring neonatal seizures (Rennie et al. 2004). However, most clinicians in NICUs opt to use amplitude integrated EEG [aEEG or cerebral function monitoring (CFMTM)] instead, because of the ease of use and minimal need for support from EEG technology and clinical neurophysiology services (Gotman 1990; Rennie et al. 2004). Since single channel aEEG often misses short, low-amplitude, or focal seizures (Eaton et al. 1994; Rennie et al. 2004), reliable automated neonatal seizure detection using continuous multi-channel EEG monitoring using 13 to 21 scalp electrodes has the potential to help clinical decision making in the NICUs and alleviate significantly the workload of the EEG interpreters.

In the literature, a few heuristic and model-based algorithms have been proposed to detect neonatal seizures. They typically consist of if-then rules and thresholds and are called heuristic because they are derived from mimicking the experts and their strategies used to detect neonatal seizures. Autocorrelation techniques (Liu et al. 1992), rhythmic discharge detection (Gotman et al. 1997), model-based EEG parameterization (Roessgen et al. 1998), modeling and complexity analysis (Celka and Colditz 2002), wave-sequence analysis (Navakatikyan et al. 2006), pseudo-periodicity analysis (Stevenson et al. 2012) and atomic decomposition (Nagaraj et al. 2014) are some of the best known methods. Furthermore, an automated neonatal seizure detector mimicking a neonatal seizure expert was proposed in our group by (Deburchgraeve et al. 2008) and was refined in (Deburchgraeve 2010). In addition, artifact removal using different blind source separation techniques has been added to the detector to improve the performance (De Vos et al. 2011). Additionally, the performance of this method has been validated on an extensive dataset of asphyxiated neonates in the NICU of the Erasmus University Medical Center (EMC) Rotterdam (Cherian et al. 2011). The total good detection rate and positive predictive value (PPV) of this method, primarily reported to be 62% and 74% respectively, improved to 84% and 90% after removing four specific patients and some dubious seizures (Cherian et al. 2011). In this paper, this method is referred to as “heuristic” algorithm.

On the other hand, machine learning approaches have also been applied to train data-driven classifiers for this problem. The following methods have been considered: time-frequency based analysis and multi-layer perceptrons (MLPs) (Hassanpour et al. 2004), quantitative features and a linear discriminant classifier (Greene et al. 2008), support vector machine (SVM) based classifier (Temko et al. 2009), Bayesian classifier via Gaussian mixture models (Temko et al., 2009), adaptive multi-channel information fusion (Li and Jeremic 2011), SVM classifier and Kalman filter (Bogaarts et al. 2014), and trend template analysis with SVM classifier tested on fetal lambs (Zwanenburg et al. 2015).

In addition, multi-stage classification composed of heuristic rules supplemented by a data-driven classifier was applied in (Aarabi et al. 2007) and (Mitra et al. 2009). In the former, a heuristic algorithm is used for artifact removal and EEG segmentation. Afterwards, the features are extracted from the segments and MLPs are applied as a classifier to identify the seizures. In the latter, conversely, MLPs

and a clustering technique are used to detect and cluster seizures (stage I, II) and then a heuristic model is applied to remove artifacts (stage III).

In this study, we describe a method for improving a previously developed automated multi-channel EEG-based neonatal seizure detector, a so-called multi-stage classifier, as explained in (Ansari et al. 2015). In the first stage, the heuristic algorithm mimicking an expert EEG reader detects the seizures. Then, in the second stage, a data-driven post-processor identifies the main characteristics of the detected segments such as evolution of spikes, synchronization of EEG and polygraphic signals, and other time-frequency domain features, in order to remove the falsely detected segments. An extensive test on three independent datasets exhibits the improved false alarm rate (FAR) in comparison to the original heuristic algorithm and its extensions.

B. Data Description and Methods

The used database composed of EEG-polygraphy recordings from 71 neonates acquired in the NICUs of Sophia Children's Hospital (part of the Erasmus University Medical Center Rotterdam, The Netherlands) (EMCR) and the NICU of the University Hospital of Leuven, Belgium (UZL). The polygraphic signals include electrocardiogram (ECG), electro-oculogram (EOG), chin or limb surface electromyogram (EMG), and abdominal respiratory movement signal (Resp.). All the neonates monitored in the EMCR (n=48) had postasphyxial HIE whereas the included neonates in the UZL (n=23) had different etiologies: HIE (n=6), metabolic (n=5), stroke (n=5), genetic (n=2), and others (n=5). During the study period, all term neonates admitted to the NICUs with presumed postasphyxial HIE or with a high clinical suspicion of seizures underwent continuous EEG (cEEG) along with video for 24-48 hours and magnetic resonance imaging (MRI). Inclusion criteria for asphyxia were either a five minute Apgar score below six or an umbilical artery pH < 7.10 and clinical encephalopathy according to Sarnat score. When seizures were detected (either electro-clinical or electrographic) treatment with anti-epileptic drugs (AED) was initiated by protocol (Cherian et al., 2011). Newborns with heart malformation were excluded. All recordings were fully anonymized in their centers. The Erasmus MC medical ethics committee approved a study (2003-2007) to assess the utility of continuous EEG monitoring in neonates with postasphyxial hypoxic ischemic encephalopathy. Use of anonymized EEG data from this study, for analysis and research was subsequently approved. Furthermore, the study was approved by the medical ethics committee of UZ Leuven.

For this retrospective study, there was no preselection of data and no EEG recordings were excluded due to low-quality of EEG recordings, artifact contaminations, or expression of dubious seizures. Definite seizures were defined as paroxysmal EEG patterns with a change from ongoing background activity with repetitive spike-trains, oscillations or a mixture there of, with clear-cut onset and offset, lasting for at least 10 seconds. The dubious seizures are paroxysmal EEG events lasting for at least 10 seconds, composed of arrhythmic mixed oscillations or sharp waves of low amplitude ($< 30\ \mu V$) with irregular variation in amplitude, frequency and morphology (without well-defined evolution) (Cherian et al. 2011). Figure 1 illustrates an arrhythmic dubious seizure with a low frequency and amplitude, with ill-defined onset and offset. In practice, the clinicians in the NICUs do not initiate treatment with anti-epileptic drugs (AED) when a dubious seizure is detected, unless this pattern frequently repeats itself or is accompanied by definite seizure patterns.

The database is partitioned into four datasets (DB1-DB4) according to their centers and durations of the scored EEG recordings. Some general characteristics of these datasets are mentioned in Table 1. DB1-DB3 were scored by a different rater, compared to DB4. DB1 has previously been used (about 8 hours for each patient) for developing the heuristic algorithm (Deburchgraeve et al. 2008). This dataset was re-used here to develop and train the proposed data-driven post-processor (using the whole EEG recordings). The rest of the datasets have not been involved in the training phase in any

way. In DB3 and DB4, only 2 hours of each recording, which had at least one seizure observed by the rater, have been selected. The data sets DB1-3 (EMCR center) were recorded at 256 Hz sampling frequency, using 17 scalp electrodes (including Cz) according to the 10-20 International System. The DB4 (UZL center) was recorded at 250 Hz using nine scalp electrodes (no F3-4, P3-4, F7-8 and T5-6), a restricted 10-20 system using 9 electrodes was used (Cherian et al. 2009). Since the heuristic method was developed only for 256 Hz, DB4 was up-sampled in advance. Then, all EEG data were filtered between 1 and 20 Hz and 20 bipolar channels were made for neonates in DB 1-3 while for the patients of DB4, 12 bipolar channels were made.

Table 1. EEG datasets

Name	Train/ Test	Center	Number of neonates	Duration (h)	Number of seizures	Total seizure burden (h)	Average seizure duration
DB1	Train	EMCR	17	461	1398	16.6	43 sec
DB2	Test	EMCR	18	489	1758	29.5	60 sec
DB3	Test	EMCR	13	27	217	4.0	66 sec
DB4	Test	UZL	23	46	120	5.6	168 sec

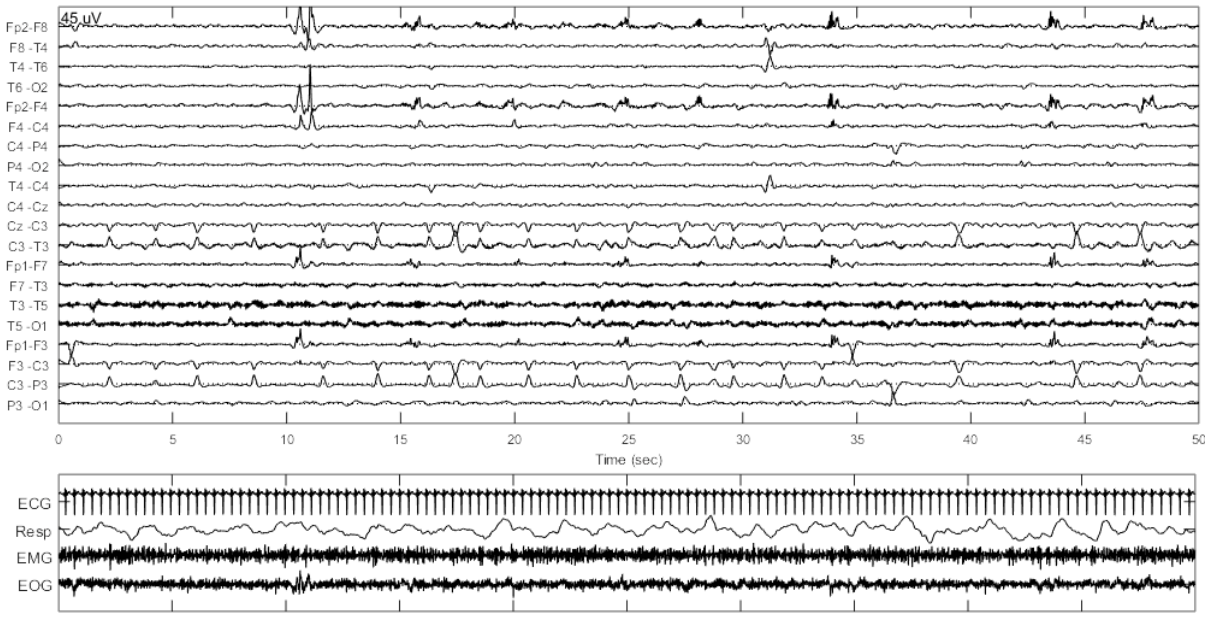

Figure 1. An arrhythmic dubious seizure observed at C_3. The frequency of spikes varies from 0.25 to 0.5 Hz and the peak-peak amplitude is lower than $20\mu V$. This dubious seizure is associated with very severe abnormality of EEG background (grade 8).

C. Proposed Method

The proposed method comprises two main stages: a heuristic classifier and a data-driven post-processor. In the first stage, the previously developed heuristic algorithm detects both the spike-train and oscillatory seizure type segments. At this stage, some artifacts or nonseizure segments having similar oscillatory or spike-train patterns (such as ECG artifacts) are incorrectly detected as seizure. In

the second stage, a pre-trained data-driven post-processor reanalyzes the heuristic detections using some informative features and a support vector machine. Figure 2 shows the schematic overview of the stages. The following subsections describe each step in detail.

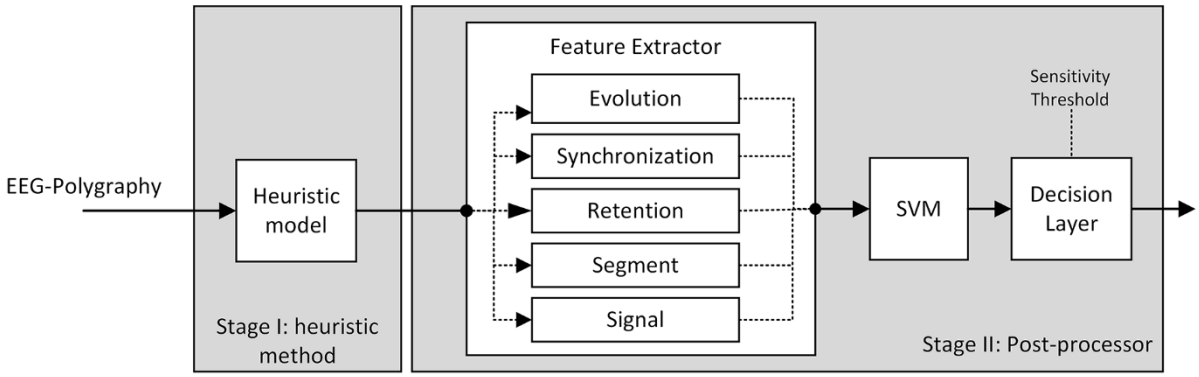

Figure 2. Diagram of the multi-stage detector. The input of the post-processor is the detected segments of the heuristic model.

a. Stage I: Heuristic Algorithm

The heuristic algorithm is comprehensively described in (Deburchgraeve et al. 2008). Briefly, it is composed of two parallel procedures; detection of spike-train and oscillatory patterns.

Spike-train detection: First, the total nonlinear energy of the signal is extracted by the Teager-Kaiser nonlinear energy operator ($TKE = x_n^2 - x_{n-1}x_{n+1}$). Then, the signal is split into 5s epochs with 80% overlap. The *"peaks"* of epochs are detected where TKE is higher than the adaptive $TKE_Threshold$. The *"spikes"* are those *"peaks"* that last more than 60 ms and are isolated from the background activity. If the overall cross-correlation of at least 6 sequential spikes is higher than 0.8, it is accepted as a "spike-train" and marked as a seizure. **Oscillatory type detection:** First, the EEG is transformed and filtered to the δ (0.5-4Hz) and θ (4-8 Hz) frequency bands using a discrete wavelet transform. Then, 3s epochs, so-called *"potential activities"*, are detected when the energy of the epoch is significantly higher than the background energy. Next, autocorrelation analysis detects the periodic activities and identifies them as oscillatory type seizures. **Mixed type:** If a segment is marked as seizure by both detectors, it is called mixed type seizure. The heuristic algorithm is a single-channel method and is applied to each channel separately. Hence, a segment can be detected as different types of seizure on different channels at the same time.

The described method is the basic algorithm published in (Deburchgraeve et al. 2008). Despite the fact that (Deburchgraeve 2010) and (De Vos et al. 2011) have improved and extended this method for decreasing the false alarm rate, in the current multi-stage method, the basic algorithm without the improvements was used for the first stage. The main reason is that although the basic algorithm detects more seizures and false alarms, the proposed post-processor (the second stage) is strong enough for removing such false alarms. Subsequently, more seizures are detected at the end and it results in a higher sensitivity when compared to the previously mentioned extensions. In the **Discussion**, the performance of the multi-stage algorithm is compared with the basic and extended versions.

b. Stage II: Data-Driven Post-Processor

The detected segments of the heuristic algorithm are the inputs of the second stage. The following steps describe the procedure.

1. All heuristic detections are split into 8s epochs with 50% overlap.
2. Five sets of features (50 features in total) are extracted for each epoch (i Feature extraction).
3. A pre-trained SVM classifier assigns a class membership probability to each epoch (0

4. **Classification**).
5. A decision layer aggregates the probabilities of all channels and classifies the detections into falsely and truly detected segments (iii Decision Making Layer).
6. The segments classified as falsely detected segments are removed and the performance is measured.

i. Feature extraction

Five sets of features (50 features in total) are extracted from each segment detected as seizure by the heuristic algorithm in the first stage. They include Evolution (12), Synchronization (4), Retention (5), Segment (5), and Signal features (24). The first four sets are extracted for the entire segment and repeated similarly in all epochs of that segment and the last set is extracted for each epoch separately.

Evolution features: The evolution of seizure patterns is one of the most characteristic features of neonatal seizure patterns as determined by visual analysis and is very useful in distinguishing it from rhythmic artifacts. In this method, three types of evolution have been taken into account: amplitude, frequency, and morphology. Because of the fact that the onset and the end of seizures have usually different (or opposite) evolutions, the segment is split into two subsections ('onset' and 'end') by determining the center of the segment. Then, evolutionary features of each subsection are extracted separately.

- **Detecting the Center:**
  - TKE of EEG signal is measured.
  - TKE is smoothed by a central linear moving average filter (MAF) (3s window).
  - The center is located where the filtered energy is maximum.

  This temporal point represents the time instant at which the overall EEG activity is maximum (Figure 3. a).

- **Evolution of Amplitude:**
  - The absolute value of the EEG signal is filtered by a MAF (0.5s window).
  - The local peaks of the filtered signal are determined.
  - A line is fitted through the peaks using robust linear regression (Holland and Welsch 1977).
  - The number of peaks is used as a feature, so-called validity of regression.
  - The slope of the fitted line is also used as evolution of amplitude (Figure 3. b).

- **Evolution of Frequency:**
  - The current subsection is split into 2s epochs with 75% overlap.
  - The power spectral density (PSD) is computed using the fast Fourier transform (FFT).
  - The mean normalized frequency (center of gravity of the PSD) is measured for each epoch.

- A line is fitted through the mean frequencies of all epochs using robust linear regression (Holland and Welsch 1977).
- The number of epochs is used as validity of regression.
- The slope of the fitted line is used as evolution of frequency (Figure 3. c).

- **Evolution of Morphology:**
  - The epoch that started 250ms before the center of the segment and lasting for 500 ms is selected as template.
  - The normalized cross-correlation is measured between the template and the current subsection.
  - The local peaks of the cross-correlation coefficients are identified.
  - A line is fitted through the peaks using robust linear regression (Holland and Welsch 1977).
  - The number of peaks is used as validity of regression.
  - The slope of the fitted line is used as evolution of morphology (Figure 3. d).

As a result, for each of the two subsections ('onset' and 'end'), 3 evolutionary slopes and 3 numbers of validity of regression are extracted. In total, every segment has 12 evolutionary features.

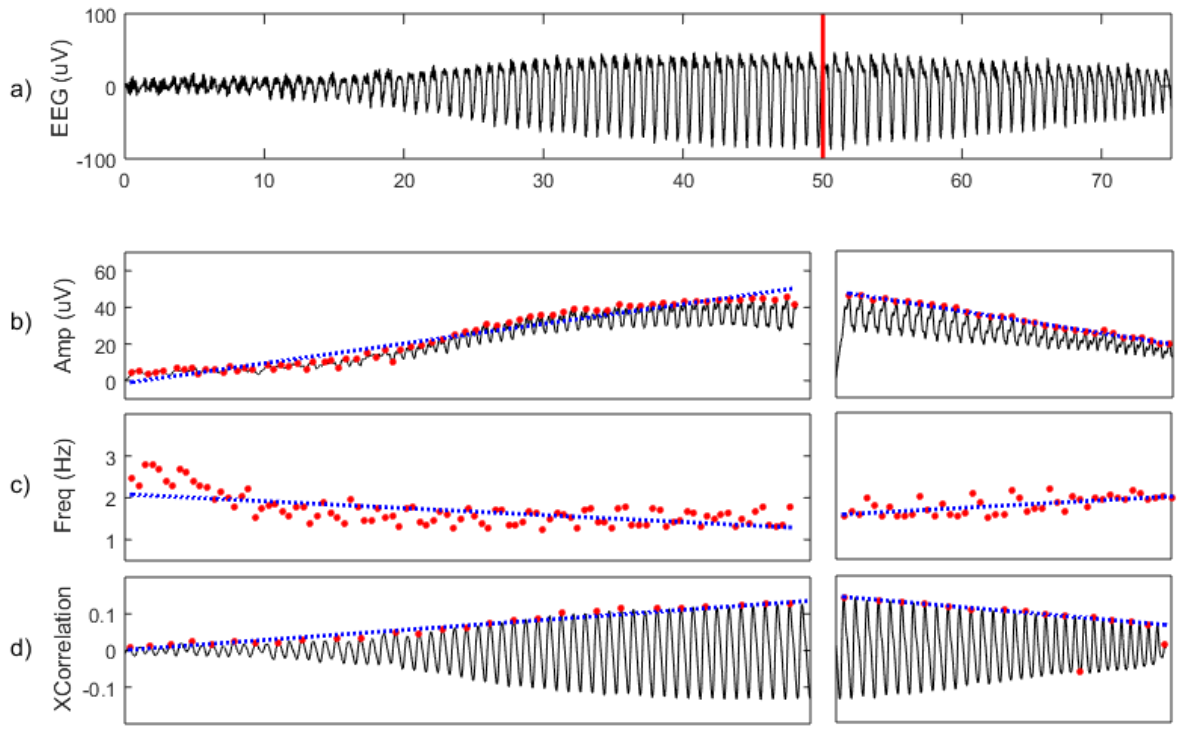

Figure 3. The behavior of the amplitude, frequency, and morphology evolution for a detected seizure. (a) Original detected EEG segment lasting for 75 seconds. The vertical red line shows the detected center of the segment where TKE is maximum. (b) Smoothed signal of rectified EEG amplitude (black line), detected local peaks of the smoothed amplitude (red dots), and the fitted line over the peaks (blue dashed line). The slope of this line determines the Amplitude Evolution. (c) The mean power frequency of 2s epochs of the EEG (red dots) and the fitted line (blue dashed line). The slope of this line determines the Frequency Evolution. (d) The normalized cross-correlation between the EEG and the template (black line), detected local peaks of the cross-correlation signal (red dots), and the fitted line (blue dashed line). The slope of this line determines the Morphology Evolution. In the traces (b-d), the reader's left side shows the 'onset' subsection and the right side shows the 'end' subsection.

Synchronization features: One of the causes of false detections of automated seizure detectors are the physiological artifacts contaminating the EEG signals such as ECG spikes, respiratory artifacts, tremor artifacts, eye movements, and blood vessel pulsations. In practice, clinical neurophysiologists often employ the recorded polygraphic signals to distinguish these kinds of artifacts during the visual assessments. In other words, they look for synchronization between the EEG spikes and polygraphic events (such as the QRS complex in the ECG signal). In order to quantify such synchronization, Mean Phase Coherence (MPC) of angular distributions is employed and calculated by:

$$MPC_{(ab)} = \left(\left[\frac{1}{N} \sum_j sin\, \phi_{ab}(j\Delta t) \right]^2 + \left[\frac{1}{N} \sum_j cos\, \phi_{ab}(j\Delta t) \right]^2 \right)^{\frac{1}{2}}, \qquad (1)$$

where N is the number of samples, $j \in [0, N-1]$, and $\phi_{ab}(t)$ is the phase difference as:

$$\phi_{ab}(t) = \phi_a(t) - \phi_b(t) = arctan \frac{\widetilde{s_a}(t) s_b(t) - s_a(t) \widetilde{s_b}(t)}{\widetilde{s_a}(t) \widetilde{s_b}(t) + s_a(t) s_b(t)} \quad (2)$$

with $s_x(t)$ and $\widetilde{s_x}(t)$ denote the signal and its Hilbert transform. The latter one is calculated by:

$$\widetilde{s_x}(t) = -i\, \mathcal{F}^{-1}\{\mathcal{F}\{s_x(t)\} sign(\omega)\}, \quad (3)$$

where $\mathcal{F}\{\,.\,\}$ and $\mathcal{F}^{-1}\{\,.\,\}$ are the Fourier transform and its inverse respectively (Mormann et al. 2000). The main advantage of using MPC is that the highly discrepant amplitudes of polygraphic signals and EEG signals would have no effect on the synchronization measurement. Moreover, MPC is a robust measurement against noise for phase synchronization of biomedical time series (Mormann et al. 2000). To illustrate the difference of MPC between an artifact and an actual seizure, Figure 4 plots a falsely detected ECG artifact (a, MPC: 0.21) and a truly detected rhythmic seizure (b, MPC: 0.04). In the current work, the MPC of EEG segment and the polygraphic signals ECG, EMG, EOG, and Resp. are measured and used as four synchronization features (Ansari et al. 2015).

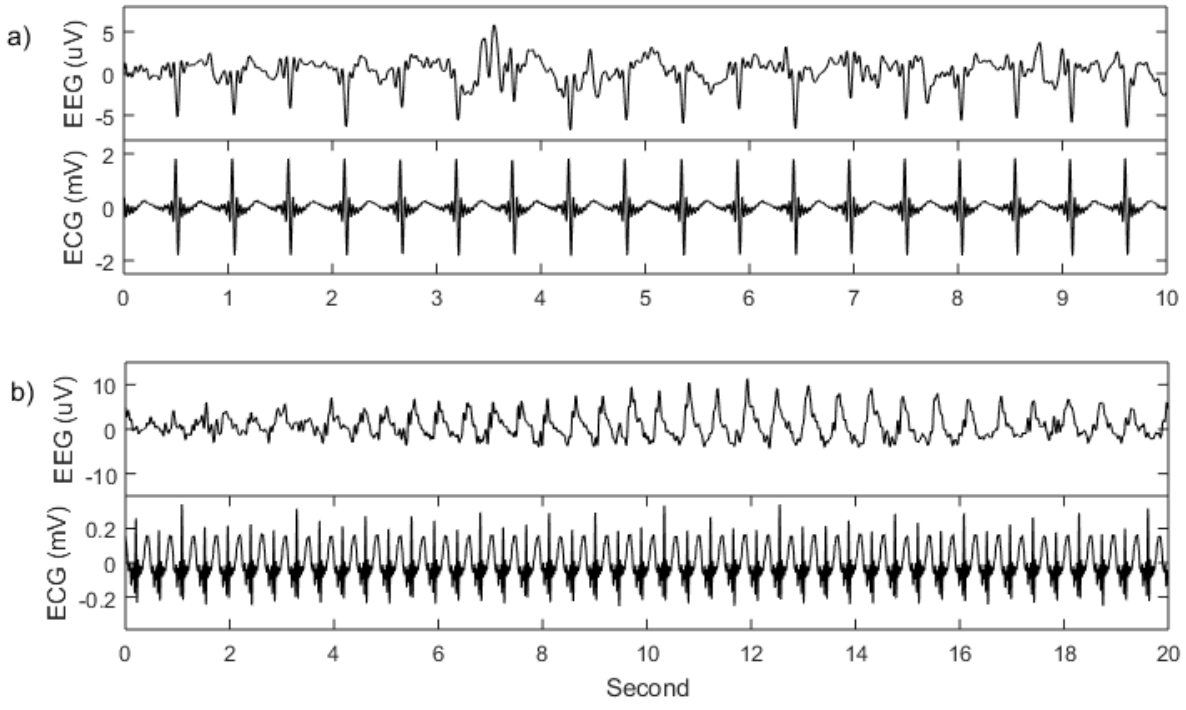

Figure 4. Two detected segments by the heuristic algorithm, (a) is a falsely detected ECG artifact ($MPC: 0.21$) and (b) is a truly detected rhythmic seizure ($MPC: 0.04$).

Retention features: In clinical practice, the physician interpreting the ongoing cEEG signal would initially appraise a number of seizure-like patterns and would consider start of AED treatment only if the patterns are repeated over time. Furthermore, seizure burden and the spatial concentration can affect the visual detection and subsequent decision to initiate treatment. In order to take this fact into account, five features are extracted from one-hour signals before the onset of every detected segment, called Retention features. They consist of (I) the number of seizures detected by the heuristic algorithm in the same channel and (II) in all channels per hour and in addition, (III) the seizure burden detected in the same channel and (IV) in all channels per hour. In order to solve the lack of data in the beginning of recordings or after missing data, a fifth feature is defined. It is 1 if at least one-hour data before the segment is available. Otherwise, it equals the duration (in hours) of the available data (so it is always between 0 and 1). In other words, the fifth feature notifies the classifier how reliable the other four Retention features are.

Segment features include 5 features extracted from the detected segment: 1) the length of the segment, 2) the number of channels expressing this seizure, 3) the type of the detection (1: spike-train, 2: oscillatory, or 3: mixed), and 4,5) (x, y) coordinates of the detected channel using a simplified two-dimensional grid on 10-20 system of electrode placement map (Ansari et al. 2015).

Signal features comprise 24 features extracted directly from the EEG epochs (8s). These features are selected from the features used in (Greene et al. 2008) and (Temko et al. 2011a) by a Lasso feature selection technique (Guyon and Elisseeff 2003) and listed in Table *2*. The formulas of the features are exhaustively described in (Greene et al. 2008).

Table 2. Selected signal features.

Type	Feature
Time Domain	• The total number of maxima and minima • Root mean square amplitude • Hjorth Complexity • Auto regression modelling error (order: 9) • Skewness • Kurtosis
Frequency Domain	• 6 spectral powers in sub-bands (1-3 Hz, 3-5 Hz, 4-6 Hz, 6-8 Hz, 9-11 Hz, 10-12 Hz) • 6 normalized spectral powers in sub-bands (1-3 Hz, 2-4 Hz, 3-5 Hz, 4-6 Hz, 6-8 Hz, 7-9 Hz) • Peak frequency • Spectral edge frequency (80%) • The wavelet energy of (1-2 Hz)
Information Theory	• Shannon entropy • Singular value decomposition entropy • Spectral entropy

ii. Classification

In order to classify the detected segments, a support vector machine (SVM) with radial basis kernel function (RBF) is used. For optimizing its hyper-parameters, leave-'one patient'-out (LOPO) cross-fold validation was applied on the training dataset and the best set of the hyper-parameters were selected. Basically, the output of the SVM for each epoch is a binary value $\{0, 1\}$ corresponding to the class labels {false detection, true detection}. However, for measuring the membership probability, a scaling method was proposed in (Platt 1999) and improved in (Lin et al. 2007). The latter one is used here to assign a probability of being a true detection to every epoch. Consequently, a detected segment which was split into 8s epochs (with 50% overlap) is mapped by the SVM and the scaling method onto a vector of probabilities. Then, the decision making layer processes this probability vector.

iii. Decision Making Layer

The task of this layer is aggregating the probabilities of epochs and channels of each detected segment. First, for aggregating the epochs of every detected channel, the vector of probabilities is compared to a threshold, called 'sensitivity threshold', and majority voting is applied to assign one label (0: 'false detection' or 1: 'true detection') to that channel. Second, for aggregating the channels, a logical "OR" operator is applied to the labels from the channels to define the truly detected segments. It means that if a detected segment has at least one channel labeled as 'True detection' (after the majority voting), it remains in the detection list. Otherwise, the segment is marked as false detection and is removed (Figure 5).

In order to choose the sensitivity threshold, first it varies from 0 (where no seizure is removed) to 1 (where all seizures are removed) and forms the Receiver Operating Characteristic (ROC) curve. Then, the elbow of the ROC showing the trade-off between the good detection rate and false alarm rate of the training data (DB1) measured by LOPO is selected (see section a).

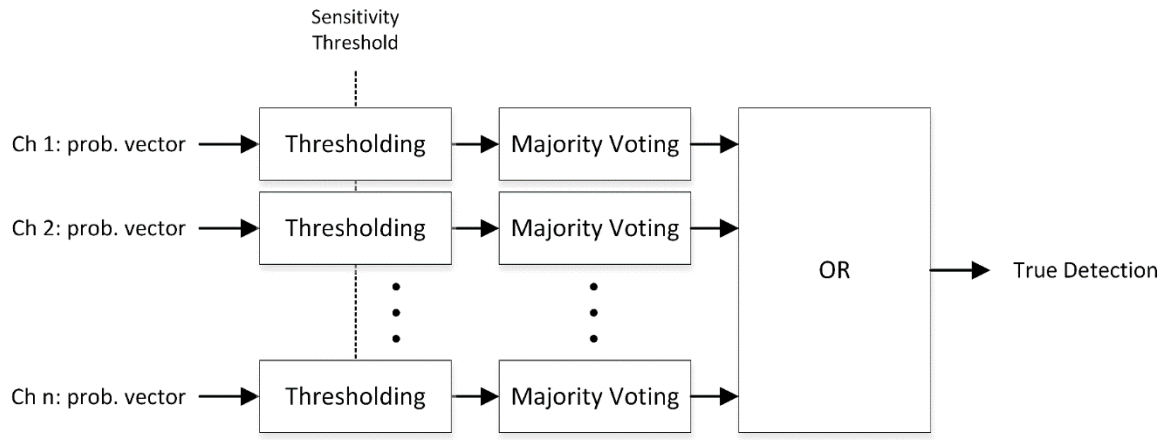

Figure 5. Diagram of the decision making layer.

D. Results

a. Improvement of performance metrics

Figure 6 shows the ROC curves for DB1-4. The curve of DB1 (training dataset) results from LOPO cross-fold validation. When the sensitivity threshold (TH) is 0, none of the detections are removed by the second stage. Therefore, the dark circles (where TH=0) are demonstrating the performance of the heuristic method (the first stage). Increasing the TH results in removing more detections and subsequently decreases the Good Detection Rate (GDR) and the False Alarm Rate (FAR). Two elbows of the curve of the training dataset (DB1) expressing the trade-offs between decrease of the GDR and FAR have been approximately chosen at 0.1 and 0.3. In this figure, the GDR and FAR of those THs for each dataset are defined by the squares and triangles respectively. Furthermore, Table 3 compares the event-based and epoch-based performance metrics of the heuristic and the proposed methods on the test datasets at these thresholds. Additionally, the averaged performance metrics measured on all patients of the test DBs are reported. It shows that the proposed post-processor (when TH=0.3) is able to decrease FAR by 64% whereas the GDR decreases 7%.

b. Analysis of event duration

Figure 7 shows the histograms of the total true and false detections of the test datasets as a function of segment duration. Comparison of the first bars of the upper and lower histograms (<30s) reveals that detecting very short seizures lasting less than 30 seconds is the most challenging task of the post-processor since it includes 26% of true and 60% of false detections. However, it removes 40% and 78% of such short true and false detections respectively and results in an increase of the Positive Predictive Value (PPV) from 20% to 40% (for detections <30s). However, it is not yet very satisfactory. The detections lasting more than one minute are almost not falsely removed by the post-processor. The analysis also shows that the duration of the seizures need to be taken into account when studies regarding seizure detection are evaluated.

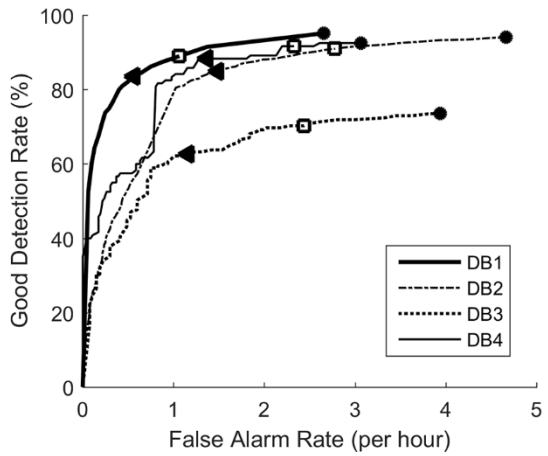

Figure 6. ROC curves of variation of the Good Detection Rate (%) against False Alarm Rate (h^{-1}) for the four datasets. The marked points indicate the performance at different sensitivity thresholds: 0 (●), 0.1 (□), and 0.3 (◀).

Table 3. The difference in performance between the heuristic and proposed methods.

Method		DB2	DB3	DB4	Average
Heuristic	GDR %	94	74	93	88
	FAR (h^{-1})	4.66	3.94	3.07	3.81
	Sensitivity %	92	61	80	79
	Specificity %	78	92	92	88
Proposed TH=0.1	GDR %	91	70	92	86
	FAR (h^{-1})	2.78	2.44	2.33	2.5
	Sensitivity %	91	60	80	79
	Specificity %	80	93	93	89
Proposed TH=0.3	GDR %	85	63	88	81
	FAR (h^{-1})	1.49	1.16	1.37	1.36
	Sensitivity %	89	58	79	77
	Specificity %	82	95	93	90

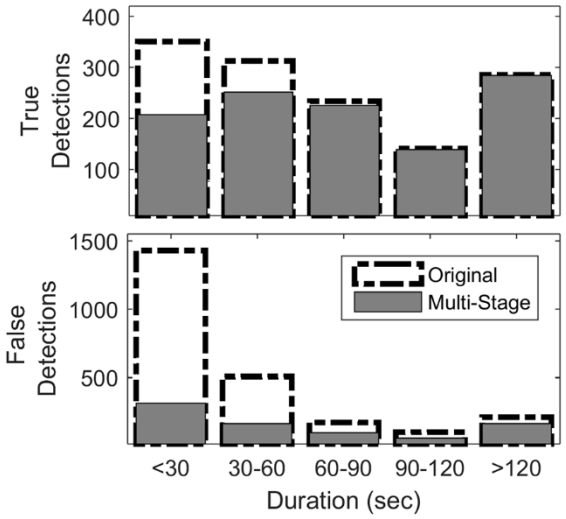

Figure 7. Histogram of the true and false detections for the original heuristic method (dashed bars) and the proposed multi-stage method (grey bars) as function of duration.

c. Analysis of EEG experts' agreement

All detections of the DB3 and DB4 were rescored by four secondary independent EEG experts and labeled as 'Definite seizure', 'Definite artifact', or 'Dubious'. Subsequently, for each detection, 5 labels are available (One primary and four secondary). Then, the majority voting is applied to assign the final label and the percentage of the majority vote is calculated. As a result, all the detections are classified into 'poorly agreed' (<60% of votes), 'moderately agreed' (60% to 80%), and 'highly agreed' (≥80%). Figure 8 shows the histogram of the definite seizures and the definite artifacts as a function of the agreement classes. The post-processor (TH = 0.3) removes only 4% of highly agreed seizures, while 66% of highly agreed nonseizure detections are removed. It increases the PPV of this class from 57% to 79% which means 22% more reliable detections of well-defined events.

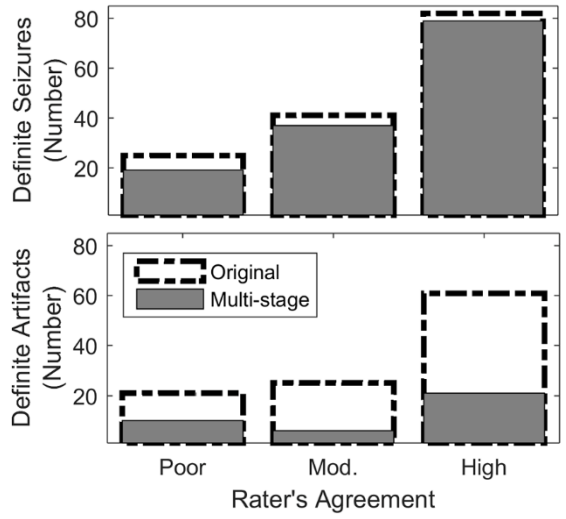

Figure 8. Histogram of the definite seizures and artifacts for the original heuristic method (dashed bars) and the proposed multi-stage method (grey bars) as a function of rater's agreement.

d. Feature Ranking

In order to rank the feature sets, two methods have been used to assign an importance score to each feature. The first method is fitting a linear ridge regression and using the absolute value of the standardized coefficients of the feature vectors (Tibshirani 1996). The second one is measuring mutual information (MI) between the features and the seizure labels (Guyon and Elisseeff 2003). Table 4 sorts the feature sets based on the averaged coefficients and mutual information values. The order of the features is similar in both methods, except for the Retention features. The reason may be high nonlinear correlation between the Retention features and the labels which is not measurable by the linear ridge method.

Table 4. Ranked feature sets using ridge regression and MI.

Rank	Ridge (Standardized Coefficients)	MI (MI value)
#1	Segment (9.00)	Retention (0.20)
#2	Evolution (4.98)	Segment (0.14)
#3	Signal (0.02)	Evolution (0.14)
#4	MPC (0.02)	Signal (0.05)
#5	Retention (0.01)	MPC (0.03)

E. Discussion

In typical pattern recognition problems such as neonatal seizure detection, the ground truth must be initially identified in order to train and test the method. In neonatal seizure detection, due to the lack of other biological markers for neonatal seizures, only visual analysis of expert clinical neurophysiologists is used and considered the 'gold standard'. Therefore, automated seizure detection (ASD) needs to be considered as an extension of the visual analysis. It can thus function as a tool for supporting clinical decision making in a busy NICU. The final decision regarding treatment initiation is best left to the team of treating physician and clinical neurophysiologist.

However, visual analysis is imperfect science considering the rather modest inter-rater agreement especially in the presence of dubious seizures. In other words, high agreement between human raters is obtained in only very typical/classical seizure patterns with well-defined morphology, reasonable amplitude and duration. Thus, if an automated seizure detector is developed solely to identify such well-defined patterns, it would mean not acknowledging the rich variety encountered in seizure patterns for practical applications. In order to build a realistic decision support tool, different types of clear and dubious patterns should be used to develop, train, and validate the algorithms. Besides, a multi-rater analysis of the automatically detected seizures can measure the overall satisfaction of different experts if they want to use that system in their NICUs. Without automated methods, sufficiently powered large multicenter cEEG studies cannot be done. There is a pressing need for such studies to address gaps in present knowledge, such as whether aggressive treatment of subtle and subclinical neonatal seizures with antiepileptic medication will improve the outcome of neonates.

In this work, different patients recorded in two centers and having different etiologies were used to train and test the proposed method and compared with the original one. Using the clear and dubious seizures at the same time provides a challenge for the methods, as would be expected in real-life scenarios. The multi-rater analysis showed the added-value of the proposed method for detecting the well-defined events.

During visual neonatal seizure detection by clinical neurophysiologists while providing clinical service to a NICU, in addition to the raw EEG data from a specific time frame, different types of information are used to identify the seizure, such as spatial distribution, corresponding compressed EEG trend signal such as the aEEG signal, the overall background activity and status of previous or subsequent seizures, evolution of patterns, etc. In addition, the EEG signal is usually being observed in frames of different duration depending on the background activity, type of seizure, morphology, EEG amplitude and other characteristics that the expert has learned over years. In ASD, in order to have a similar

approach and extract some of this extra information, the length of the processing window should be variable. For instance, a short processing window (a few seconds) is required to extract fast-rate features like frequency information, while a longer window (about a minute) is needed for considering the evolution of patterns, or very long window (about hours) should be used to take the features of the previous seizures into account.

In this study, unlike most retrospectively proposed automatic seizure detectors having a (relatively) fixed-size window, different window lengths varying from 8 seconds to 1 hour are applied to extract the 'core characteristics' of seizures including: the features of the EEG signal at a specific time (signal features), the spatial information (segment features), the evolution of patterns (evolution features), the correlation of EEG and polygraphic signals (synchronization features), the seizure burden and repetition detected in the previous one hour (Retention features). As a result, the FAR drops by 64% while the GDR decreases by 7% when the TH is 0.3, or the FAR decreases by 32% with only 2% reduction of the GDR when TH equals 0.1.

However, It should be taken into consideration that the described 'core characteristics' of seizures in this article are purely based on visual patterns and not based on any pathophysiological basis for these patterns. Without more basic research into underlying seizure mechanisms at cellular, synaptic and network levels, the processes underlying these characteristics cannot be elucidated. Other modifying factors such as severity and nature of underlying brain injury and treatment with antiepileptic medications also need to be taken into account.

Table 5 lists the performance of the heuristic algorithm and its extensions compared to the proposed multi-stage classifier. It shows that, the refined version of the heuristic algorithm (Deburchgraeve 2010) results in a GDR of 67% with 1.7 FAR per hour on the test DBs on average per patient. However, by keeping the same FAR (corresponding to TH=0.23) the multi-stage method produces a GDR of 82%. Besides, if the blind source separation (BSS) technique (De Vos et al. 2011) is applied on the EEG signal in order to pre-process and remove the polygraphic-related artifacts the FAR decreases to 1.2 (h^{-1}) with a GDR of 66%. However, the multi-stage method (without pre-processing) has a 14% higher GDR for the same FAR (where TH = 0.33). Additionally, because the multi-stage method has less time complexity (in recall) compared to BSS, it post-processed the test DBs about 3 times faster than the pre-processor which is an important advantage for real-time implementation.

Table 5: Performance of the original heuristic algorithm and its extensions compared to the proposed method.

Method	DB2		DB3		DB4		Average	
	GDR	FAR	GDR	FAR	GDR	FAR	GDR	FAR
(Deburchgraeve et al. 2008) = Multi-stage (TH = 0)	94%	4.6	74%	3.9	93%	3.1	**88%**	**3.8**
(Deburchgraeve 2010)	76%	2.6	49%	2.1	72%	0.9	67%	1.7
Multi-stage (TH = 0.23)	87%	1.7	65%	1.6	88%	1.8	**82%**	**1.7**
(De Vos et al. 2011)	69%	1.4	49%	0.8	73%	1.5	66%	1.2
Multi-stage (TH = 0.33)	84%	1.4	62%	1	88%	1.3	**80%**	**1.2**

Figure 9 shows the performance of the proposed method compared to the other reported ones. In this figure, the grey circles show the reported performance of the heuristic algorithms and its extensions in the original studies. However, the dark circles are corresponding to the grey ones when those algorithms are again applied on the test DBs used in this paper. Furthermore, the white diamond shows the performance of the algorithm that was proposed in (Temko et al. 2011a) while the dark diamond displays its performance when the algorithm is re-implemented in our group and tested in DB2-4. The considerable difference between the performance of the dark and light symbols demonstrates that comparing different algorithms while using different databases results in a very inaccurate judgment. There are many database-related factors that may lead to such inaccuracy like the difference in the number of patients, the density (prevalence) of seizures, the length of recordings, the distribution of seizure durations, and the grade of the background activities (Cherian et al. 2011; Temko et al. 2011b; Stevenson et al. 2012). The most important reason that makes our test DBs more challenging is that the dubious seizures were not removed in our datasets. Generally, detecting the dubious seizures which are usually shorter and of a very low-amplitude is a difficult task and needs a fine seizure-sensitive detector which would produce many false alarms in return. Various seizure-related factors such as severe brain injury as well as aggressive treatment with AEDs such as midazolam and lidocaine might influence seizure morphology and rhythmicity and lead to more dubious patterns (Cherian et al., 2011). The automated seizure detection algorithm would help tackle this kind of problems.

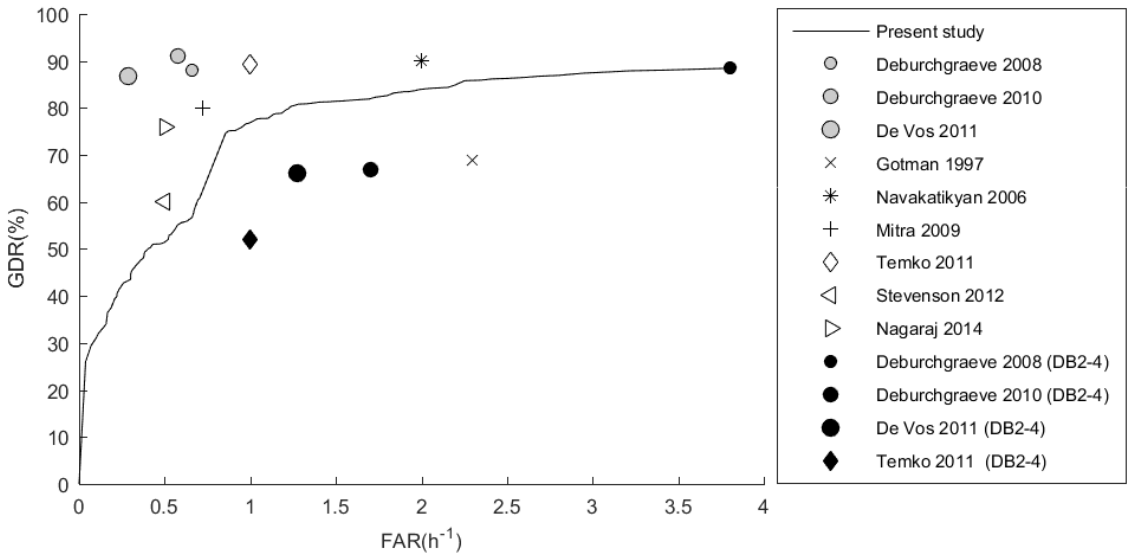

Figure 9. Performance of the proposed multi-stage classifier tested on DB2-4 compared to the reported performance of other methods. The grey circles show the previously reported performance of the heuristic method tested on different databases. The dark circles show the performance of the heuristic method and its extensions when it is applied on DB2-4. Furthermore, the light diamond shows the GDR (with one false alarm per hour) reported in (Temko et al. 2011a) while the dark diamond shows the GDR of its re-implementation tested on DB2-4.

F. Conclusions

A major improvement of a previously developed automated neonatal seizure detector was achieved by combining a machine learning technique with the heuristic algorithm. In such a multi-stage detector, first a simplified visual seizure detection approach of an EEG expert is formulated by a few heuristic if-then rules and thresholds. Then, the core characteristics of seizure patterns are extracted by five sets of features including I) "evolution features" for quantifying the evolution of amplitude, frequency and morphology of detections, II) "synchronization features" for extracting the correlation of detected spikes of the EEG with those of the polygraphic signals, III) "Retention features" for measuring the seizure burden and repetition occurring before the detections, IV) "Segment features" for using general information of the detections such as spatial spread or type of seizure, and V) "signal features" for extracting the common signal processing features. Next, the features are fed into an SVM in order to distinguish between the truly and falsely detected segments and minimize classification error. The extracted features are based either on the underlying behavior of the EEG seizure patterns such as evolution, retention, synchronization etc., or their mathematical characteristics such as kurtosis, Hjorth complexity etc. The core features of neonatal seizure patterns extracted in this study through a data driven approach are suitable for researchers developing automated seizure detection methods for further testing on larger databases. We have shown that a combination of the heuristic and data-driven approaches yields a significantly improved performance.

Future developments: incorporation into neonatal specific EEG system.

Acknowledgements

AA, VM, and SV are supported by: Bijzonder Onderzoeksfonds KU Leuven (BOF): Center of Excellence (CoE) PFV/10/002 (OPTEC); Fonds voor Wetenschappelijk Onderzoek-Vlaanderen (FWO) projects: G.0427.10N (Integrated EEG-fMRI), G.0108.11 (Compressed Sensing) G.0869.12N (Tumor imaging) G.0A5513N (Deep brain stimulation); Agentschap voor Innovatie door Wetenschap en Technologie (IWT) projects: TBM 080658-MRI (EEG-fMRI), TBM 110697-NeoGuard; iMinds Medical Information Technologies: Dotatie-Strategisch basis onderzoek (SBO-2015), ICON: NXT_Sleep; Belgian Federal Science Policy Office: IUAP P7/19/ (DYSCO, 'Dynamical systems, control and optimization', 2012-2017); Belgian Foreign Affairs-Development Cooperation: VLIR UOS programs (2013-2019); European Union's Seventh Framework Programme (FP7/2007-2013): EU MC ITN TRANSACT 2012, #316679, ERASMUS EQR: Community service engineer, #539642-LLP-1-2013; EU INTERREG IVB NWE programme #RECAP 209G; European Research Council (ERC) Advanced Grant, #339804 BIOTENSORS This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information. AD is also supported by IWT PHD grant: TBM 110697-NeoGuard.

References

Aarabi A, Grebe R, Wallois F. A multistage knowledge-based system for EEG seizure detection in newborn infants. Clin Neurophysiol. 2007 Dec;118(12):2781–97.

Ansari AH, Matic V, De Vos M, Naulaers G, Cherian PJ, Van Huffel S. Improvement of an automated neonatal seizure detector using a post-processing technique. 2015 37th Annu Int Conf IEEE Eng Med Biol Soc EMBC. 2015. p. 5859–62.

Bogaarts JG, Gommer ED, Hilkman DMW, Kranen-Mastenbroek VHJM van, Reulen JPH. EEG Feature Pre-processing for Neonatal Epileptic Seizure Detection. Ann Biomed Eng. 2014 Aug 15;42(11):2360–8.

Bye AME, Flanagan D. Spatial and Temporal Characteristics of Neonatal Seizures. Epilepsia. 1995 Oct 1;36(10):1009–16.

Celka P, Colditz P. A computer-aided detection of EEG seizures in infants: a singular-spectrum approach and performance comparison. IEEE Trans Biomed Eng. 2002 May;49(5):455–62.

Cherian PJ, Deburchgraeve W, Swarte RM, De Vos M, Govaert P, Van Huffel S, et al. Validation of a new automated neonatal seizure detection system: A clinician's perspective. Clin Neurophysiol. 2011 Aug;122(8):1490–9.

Cherian PJ, Swarte RM, Visser GH. Technical standards for recording and interpretation of neonatal electroencephalogram in clinical practice. Ann Indian Acad Neurol. 2009;12(1):58–70.

Connell J, Oozeer R, de Vries L, Dubowitz LM, Dubowitz V. Clinical and EEG response to anticonvulsants in neonatal seizures. Arch Dis Child. 1989 Apr;64(4 Spec No):459–64.

Deburchgraeve W. Development of an automated neonatal EEG seizure monitor [PhD Thesis]. [Leuven]: Katholieke Universiteit of Leuven; 2010.

Deburchgraeve W, Cherian PJ, De Vos M, Swarte RM, Blok JH, Visser GH, et al. Automated neonatal seizure detection mimicking a human observer reading EEG. Clin Neurophysiol. 2008 Nov;119(11):2447–54.

Eaton DM, Toet M, Livingston J, Smith I, Levene M. Evaluation of the Cerebro Trac 2500 for monitoring of cerebral function in the neonatal intensive care. Neuropediatrics. 1994 Jun;25(3):122–8.

Gotman J. Automatic seizure detection: improvements and evaluation. Electroencephalogr Clin Neurophysiol. 1990 Oct;76(4):317–24.

Gotman J, Flanagan D, Zhang J, Rosenblatt B. Automatic seizure detection in the newborn: methods and initial evaluation. Electroencephalogr Clin Neurophysiol. 1997 Sep;103(3):356–62.

Greene BR, Faul S, Marnane WP, Lightbody G, Korotchikova I, Boylan GB. A comparison of quantitative EEG features for neonatal seizure detection. Clin Neurophysiol. 2008 Jun;119(6):1248–61.

Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. J Mach Learn Res. 2003 Mar;3:1157–82.

Hahn JS, Olson DM. Etiology of Neonatal Seizures. NeoReviews. 2004 Aug 1;5(8):e327–35.

Hassanpour H, Mesbah M, Boashash B. Time–frequency based newborn EEG seizure detection using low and high frequency signatures. Physiol Meas. 2004 Aug 1;25(4):935.

Holland PW, Welsch RE. Robust regression using iteratively reweighted least-squares. Commun Stat - Theory Methods. 1977 Jan 1;6(9):813–27.

Li H, Jeremic A. Neonatal seizure detection using blind distributed detection with correlated decisions. 2011 Annu Int Conf IEEE Eng Med Biol Soc. 2011. p. 6580–4.

Lin H-T, Lin C-J, Weng RC. A note on Platt's probabilistic outputs for support vector machines. Mach Learn. 2007 Aug 8;68(3):267–76.

Liu A, Hahn JS, Heldt GP, Coen RW. Detection of neonatal seizures through computerized EEG analysis. Electroencephalogr Clin Neurophysiol. 1992 Jan;82(1):30–7.

Mitra J, Glover JR, Ktonas PY, Thitai Kumar A, Mukherjee A, Karayiannis NB, et al. A Multi-stage System for the Automated Detection of Epileptic Seizures in Neonatal EEG. J Clin Neurophysiol Off Publ Am Electroencephalogr Soc. 2009 Aug;26(4):218–26.

Mormann F, Lehnertz K, David P, E. Elger C. Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients. Phys Nonlinear Phenom. 2000 Oct 1;144(3–4):358–69.

Murray DM, Boylan GB, Ali I, Ryan CA, Murphy BP, Connolly S. Defining the gap between electrographic seizure burden, clinical expression and staff recognition of neonatal seizures. Arch Dis Child - Fetal Neonatal Ed. 2008 May 1;93(3):F187–91.

Nagaraj SB, Stevenson NJ, Marnane WP, Boylan GB, Lightbody G. Neonatal Seizure Detection Using Atomic Decomposition With a Novel Dictionary. IEEE Trans Biomed Eng. 2014 Nov;61(11):2724–32.

Navakatikyan MA, Colditz PB, Burke CJ, Inder TE, Richmond J, Williams CE. Seizure detection algorithm for neonates based on wave-sequence analysis. Clin Neurophysiol. 2006 Jun;117(6):1190–203.

Platt JC. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Adv LARGE MARGIN Classif. MIT Press; 1999. p. 61–74.

Rennie JM, Chorley G, Boylan GB, Pressler R, Nguyen Y, Hooper R. Non-expert use of the cerebral function monitor for neonatal seizure detection. Arch Dis Child-Fetal Neonatal Ed. 2004;89(1):F37–40.

Roessgen M, Zoubir A, Boashash B. Seizure detection of newborn EEG using a model-based approach. IEEE Trans Biomed Eng. 1998 Jun;45(6):673–85.

Scher MS, Alvin J, Gaus L, Minnigh B, Painter MJ. Uncoupling of EEG-clinical neonatal seizures after antiepileptic drug use. Pediatr Neurol. 2003 Apr;28(4):277–80.

Stevenson NJ, O'Toole JM, Rankine LJ, Boylan GB, Boashash B. A nonparametric feature for neonatal EEG seizure detection based on a representation of pseudo-periodicity. Med Eng Phys. 2012 May;34(4):437–46.

Temko A, Thomas E, Boylan G, Marnane W, Lightbody G. An SVM-based system and its performance for detection of seizures in neonates. Annu Int Conf IEEE Eng Med Biol Soc 2009 EMBC 2009. 2009. p. 2643–6.

Temko A, Thomas E, Marnane W, Lightbody G, Boylan G. EEG-based neonatal seizure detection with Support Vector Machines. Clin Neurophysiol. 2011a Mar;122(3):464–73.

Temko A, Thomas E, Marnane W, Lightbody G, Boylan GB. Performance assessment for EEG-based neonatal seizure detectors. Clin Neurophysiol. 2011b Mar;122(3):474–82.

Tibshirani R. Regression Shrinkage and Selection via the Lasso. J R Stat Soc Ser B Methodol. 1996;58(1):267–88.

Volpe JJ. Neurology of the Newborn. 5th ed. Philadelphia: Saunder WB; 2008.

De Vos M, Deburchgraeve W, Cherian PJ, Matic V, Swarte RM, Govaert P, et al. Automated artifact removal as preprocessing refines neonatal seizure detection. Clin Neurophysiol. 2011 Dec;122(12):2345–54.

Zwanenburg A, Andriessen P, Jellema RK, Niemarkt HJ, Wolfs TG, Kramer BW, et al. Using trend templates in a neonatal seizure algorithm improves detection of short seizures in a foetal ovine model. Physiol Meas. 2015;36(3):369.