Published in final edited form as:
Int J Neural Syst. 2013 August ; 23(4): 1350018. doi:10.1142/S0129065713500184.

ROBUST NEONATAL EEG SEIZURE DETECTION THROUGH ADAPTIVE BACKGROUND MODELLING

ANDRIY TEMKO,
Neonatal Brain Research Group, Department of Electrical and Electronic Engineering, University College Cork, Ireland atemko@ucc.ie

GERALDINE BOYLAN,
Neonatal Brain Research Group, Department of Paediatrics and Child Health, University College Cork, Ireland g.boylan@ucc.ie

WILLIAM MARNANE, and
Neonatal Brain Research Group, Department of Electrical and Electronic Engineering, University College Cork, Ireland l.marnane@ucc.ie

GORDON LIGHTBODY
Neonatal Brain Research Group, Department of Electrical and Electronic Engineering, University College Cork, Ireland g.lightbody@ucc.ie

Abstract

Adaptive probabilistic modelling of the EEG background is proposed for seizure detection in neonates with hypoxic ischemic encephalopathy. The decision is made based on the temporal derivative of the seizure probability with respect to the adaptively modeled level of background activity. The robustness of the system to long duration 'seizure-like' artifacts, in particular those due to respiration, is improved. The system was developed using statistical leave-one-patient-out performance assessment, on a large clinical dataset, comprising 38 patients of 1479 hours total duration. The developed technique was then validated by a single test on a separate totally unseen randomized prospective dataset of 51 neonates totaling 2540 hours of duration. By exploiting the proposed adaptation, the ROC area is increased from 93.4% to 96.1% (41% relative improvement). The number of false detections per hour is decreased from 0.42 to 0.24, while maintaining the correct detection of seizure burden at 70%. These results on the unseen data were predicted from the rigorous leave-one-patient-out validation and confirm the validity of our algorithm development process.

Keywords

Neonatal Seizure Detection; EEG Background

1. Introduction

Dealing with artifacts represents the greatest challenge for automated detection of neonatal seizures[3,4]. In particular the respiration artifact is known to be a major source of error in the modern seizure detection system for neonates[1,2]. Body movements during respiration manifest as a slow rhythmic waves in the EEG. Sick newborn babies may have high

Europe PMC Funders Author Manuscripts

respiratory rates of up to 100/min and therefore some rhythmic delta activity (0–4Hz) seen in the EEG may be respiration artifact. The rhythmic nature of these artifacts mimics abnormal cerebral EEG activity. If additional respiration monitoring is not used during EEG recording, this activity could be mistaken for seizure even by an experienced neurophysiologist. Examples of seizure and respiration artifact activity are presented in Fig. 1.

A number of methods have been proposed previously in an attempt to automatically detect neonatal seizures[5–8], however to date their transition to clinical use has been limited due to poor performance. Navakatikyan *et al*.,[7] reported that their system correctly detected 82.8% of the seizure burden (the total amount of time the newborn spends in seizure) at a cost of two detections per hour (FD/h). A recent study by Cherian *et al*.,[1] reported the correct detection of on average 59% of seizure burden at a cost of 0.58 FD/h. With the exclusion of the four most difficult and worst performing patients, the number of FD/h was further improved from 0.58 to 0.28. It is frustrating to see that the performance reported in laboratory conditions[10] is not usually confirmed on a separate validation dataset[1].

There are two key directions in automated neonatal seizure detection. The first follows analytical learning principles[9] and focuses on the creation of a set of heuristic rules and thresholds from clinical prior knowledge[1,5-7]. The resultant detectors analyze EEG using a small number of the descriptors from which a decision is made using empirically derived thresholds. The second approach relies on inductive learning[9] and utilizes statistical classifier based methods[2,8,34,35], which employ elements of machine learning to classify a set of features using a data-driven decision rule.

In automated systems, when simultaneous recording of the respiratory trace is present, it can be directly correlated with the EEG signal to establish whether or not an alarmed event is due to respiration artifact[5] Alternatively, more complex independent component analysis (ICA) based decomposition with the given polygraph references can be employed[1,10]. Even then, respiration artifact cannot be completely isolated and heavily affected patients are sometimes removed from the system validation results[1]. When respiration is not recorded with EEG, which is a common situation in the NICU, then the resulting high number of false detections becomes truly challenging. A few attempts have been made to eliminate or detect respiration artifacts in EEG alone. Chervin *et al*.,[11] introduced several features that were intended to capture respiration-specific cycle changes in the EEG of a 6-year old child. Zhang *et al*.,[12] used power and relative wavelet energy parameters to characterize respiration artifacts in a rabbit EEG model. Similarly, many other methods (see Ref. 13 and references therein) only achieved reasonably good performance in the absence of seizure EEG activity.

In this work, a modification to our previously validated system presented in Ref. 2 is proposed in order to improve performance in the presence of seizure-like, long-lasting artifacts, in particular the respiration artifact. A way to estimate how much the developed seizure detector algorithm is biased by non-seizure activity present in the ongoing EEG is developed. Instead of making a decision based on the classifier probability of the seizure, the decision is made based on its temporal derivative with respect to the level of background probability which is adaptively modeled over the past probabilistic activity.

The paper is organized as follows: Section 2 briefly describes the dataset, the neonatal seizure detector previously developed by the group, and discusses the performance assessment routines used. The adaptive modelling of background level and its integration into the probabilistic framework of the developed detection is given in Section 3. Section 4 presents the comparison results. Conclusions are drawn in Section 5.

2. Neonatal Seizure Detectors

2.1. *Dataset*

The dataset in our work is composed of EEG recordings from 38 newborns recruited from the NICU, Cork University Maternity Hospital (CUMH), Cork, Ireland. The patients were full term babies ranging in gestational age from 39 to 42 weeks. A Carefusion NicOne video EEG monitor was used to record multi-channel EEG at 256Hz using the 10-20 system of electrode placement, modified for neonates. The standard protocol for EEG recording in the NICU required the following 9 active electrodes: T4, T3, O1, O2, F4, F3, C4, C3, and Cz. Then, the following 8 EEG bipolar pairs were used to annotate the data: F4-C4, C4-O2, F3-C3, C3-O1, T4-C4, C4-Cz, Cz-C3 and C3 - T3. Eighteen newborns had seizures secondary to HIE. All electrographic seizures were annotated independently by two experienced neonatal electro-encephalographers using simultaneous video EEG. All disagreements in annotations were resolved by consensus. The combined length of the EEG recordings of seizure patients totaled 816.7 hours with per patient mean/median length of 45.4/48.5 hours and contained 1389 electrographic seizures. The distribution of seizures in our dataset tightly follows the distribution of neonatal seizures in other studies[37], with ~26% of seizure shorter than 60s and ~43% shorter than 90s. The dataset contains a wide variety of seizure types including both electrographic-only and electro-clinical seizures of focal, multi-focal and generalized types.

There were no seizures observed in the remaining 20 HIE patients. The combined length of the EEG recordings of non-seizure patients totaled 662.1 hours with per patient mean/median length of 33.1/22.7 hours. The continuous EEG recordings were not edited to remove the large variety of artifacts and poorly conditioned signals that are commonly encountered in the real-world NICU environment. Therefore this dataset is truly representative of the real-life situation in the NICU and it allows the most robust estimate of the algorithm's performance. The respiration trace was not present in 15 out of 18 patients. The dataset of seizure patients used are detailed in Table 1.

Additionally, the developed algorithm is tested on a separate randomized dataset of 51 babies (24 with seizures and 27 without seizures), totaling 2540 hours in duration and containing 1142 seizures. The data were collected at CUMH between mid-2009 and mid-2011 as a part of the ongoing validation campaign for regulatory approval (FDA). These babies have mixed etiologies (asphyxia, HIE, stroke, meningitis) and have been annotated by a different clinical neurophysiologist from University College London Hospital.

2.2. *Automated seizure detection system architecture*

The neonatal seizure detection system is shown in Fig. 2. The EEG from the 8 channels was down-sampled from 256Hz to 32Hz with an anti-aliasing filter set at 12.8Hz. The EEG was then split into 8s epochs with 50% overlap between epochs.

Fifty-five features were extracted from each channel which represent both time and frequency domain characteristics as well as information theory based parameters. These features have partly been used for EEG description in both neonatal[2,8,16,17,41,42] and adult[31,32,33,35,38,39,40,41] population. The features are listed in Table 2. The results of feature combination, ranking and selection for neonatal seizure detection are not presented in this paper, but can be found in our separate study[16]. Several other system architecture choices are detailed in Ref. 2 and Ref. 17.

The training dataset consists of approximately 20 minutes of seizure EEG per-patient annotated per channel. At the same time, 40k epochs were randomly selected from the non-

seizure data of seizure patients for representation of the non-seizure class. Non seizure patients' data were not used for training at any time.

The training data for the classifier were normalized by subtracting the mean and dividing by standard deviation of each feature to assure commensurability of the various features. This normalizing template was then applied to the testing data. The normalized features extracted from each epoch were then fed to a classifier. The support vector machine (SVM) classifier with a Gaussian kernel was implemented.

In the testing stage, the output of the SVM was converted to a probability-like value using the method of Platt described in Ref. 19.

The probabilistic output was then time-wise smoothed with a moving average filter (MAF) of 15 epochs (~1 minute). The maximum of the averaged probabilities across all channels was computed to represent the final support of a seizure. It was then compared to a threshold from the interval [0 1]. After comparison, a binary decision was taken: 1 for seizure and 0 for non-seizure. The 'collar' technique was applied last – every seizure decision was extended from either side by 7 epochs to account for the delay introduced by the MAF.

2.3. *Performance assessment and metrics*

Leave-one-out performance assessment—In clinical practice, samples of testing patient data are never available beforehand in the NICU. It is therefore necessary to develop a patient-independent neonatal seizure detector. For this reason, the leave-one-patient-out (LOO) cross-validation method was used to assess the performance of the system for patient-independent seizure detection[2]. In this manner, all but one patients' data from the seizure dataset (Table 1) were used for training and data from the remaining seizure patient's along with all non-seizure patients were used for testing. This procedure was repeated until each seizure patient had been a test subject and the mean result was reported. The LOO method is known to be an almost unbiased estimation of the true generalization error[20]. What is examined with the LOO procedure is not a particular model, but indeed the methodology used to obtain such a model. This last point means that the LOO estimate effectively gives a robust prediction of the performance that other researchers or practitioners will obtain using this method, but trained on their data. Here, eighteen 17 vs. 1 data splits made by the LOO method formed the performance assessment routine.

In each of these 18 splits, nested cross-validation model selection on the training 17 patients' data was performed to choose suitable model parameters. On average, 19% of training data were support vectors. The most frequent pair of selected hyper-parameters was 0.05 for gamma in the Gaussian kernel and 20 for the generalization parameter *C*. The model selection routine is completely independent of the performance assessment routine and the testing subject was not seen or used at any time for any system parameter tuning.

Prospective validation—For validation of the system on the prospective dataset of 51 newborns, the system was trained on all 18 seizure newborns from the LOO dataset. The gamma = 0.05 and C = 20 were used to train the SVM model. All the system parameters were fixed and a single test over the validation dataset was performed. The validation dataset is truly unseen and the engineering team of the group had no access to the annotations at any time.

Metrics—The main metric used in this work is the area under the Receiver Operating Characteristic (ROC) curve which plots sensitivity versus specificity (or 1-specificity) values. Sensitivity and specificity are defined as the epoch-wise accuracy of each class (seizure and non-seizure), respectively. It is worth noting that sensitivity corresponds to the

Europe PMC Funders Author Manuscripts

percentage of the correctly detected seizure burden, i.e. the total amount of time the baby spends in seizure. While seizure burden is the most important metric to indicate whether a patient should be treated or not[4], it remains largely unrecognized[29] when using conventional event-based GDR and FD/h metrics.

A neonate might be monitored for up to 72 hours in the NICU, thus to be clinically useful, the system should work at very high specificity to produce a tolerably low number of false detections per hour (typically <<1FD/h). One can not argue with the fact that the higher the epoch-based specificity is the lower the number of FD/h will become. Thus, the regions of the ROC area where specificity is higher than 90%, 95% or 99% are of a particular importance. In this work, we report the ROC area where specificity is higher than 90% as ROC90. We also report the number of FD/h. By thresholding the probability of seizure (in the range from 0 to 1), it is possible to report the curves of performance in contrast to reporting a performance for a single operating point.

The ROC area is related to the Wilcoxon test of significance[21–23]. This relationship can be used to derive statistical properties of the ROC area such as its standard error (SE)[22]:

$$SE(\gamma) = \sqrt{\frac{\gamma(1-\gamma) + (n_A - 1)(Q_1 - \gamma^2) + (n_N - 1)(Q_2 - \gamma^2)}{n_A n_N}} \quad (3)$$

where γ is the ROC area, $Q_1 = \gamma/(2 - \gamma)$ and $Q_2 = 2\gamma^2/(1 + \gamma)$, n_A and n_N are the numbers of seizure (abnormal) and non-seizure (normal) epochs. To calculate the statistical significance of a difference of two algorithms (ROC areas) evaluated on the same data, we compute the z statistic by taking into account the correlation of the two ROC curves[23]:

$$z = \frac{\gamma_1 - \gamma_2}{\sqrt{SE(\gamma_1)^2 + SE(\gamma_2)^2 - 2rSE(\gamma_1)SE(\gamma_2)}} \quad (4)$$

where γ_1 and γ_2 refer to the observed areas associated with algorithm 1 and 2, respectively. Here, r represents the estimated correlation between the two ROC curves as outlined in Ref. 23. The resultant p values of the two-tailed test are reported and values less than 0.01 are considered significant.

3. Adaptive Modelling of EEG Background

3.1. *'Seizure like' artifacts*

Different seizure detection algorithms can suffer from different artifact types; however seizure-like artifacts represent a formidable challenge for most neonatal seizure detectors. It has been identified in our previous work[17] that respiration artifact is responsible for a number of false detections in our algorithm. To show the complexity of the problem in the absence of respiration monitoring, the respiration artifact is contrasted to seizure in the signal, spectral and probabilistic domains using Fig. 1, Fig. 3a and Fig. 3b. When seizure activity is present, the spectral parameters alone are incapable of discriminating between artifact and seizure. This is illustrated in Fig. 3 (a) where cluster centers of the log spectra are plotted for both the respiration artifact and the seizure. The log spectra are obtained by applying overlapping triangular filter-banks centered at frequencies 1-11 Hz as shown along the bottom of Fig. 3 (a). It can be seen that the spectral envelope signatures of both classes look similar. Fig. 3 (b) shows the histogram of the classifier probabilistic outputs for respiration artifact EEG and neonatal seizure EEG obtained by the detector presented in Ref. 2 and described above. The detector along with spectral envelope parameters uses a set of other EEG features (Table 2). The probability density functions are not identical, though

there is a substantial overlap produced for these two classes. It is also clear that respiration artifact results in an elevated probability. As an example the output of the moving average filter is plotted in Fig. 3 (c) for ~2 hours of EEG from patient 4. The EEG in the middle (between ~400 and ~800 epochs) is corrupted with respiration artifact.

In contrast to a respiration artifact, the seizure morphology evolves both temporally and spatially with similar evolution observed in frequency. Seizures have a definite beginning middle and end, unlike a respiration artifact. Walls-Esquivel *et al*.,[24] concluded that in the absence of a respiratory trace, observing the spatiotemporal history or context of a suspected event is essential for visual discrimination between respiration artifact and seizure in clinical practice. This approach involves the comparison of the current quantized EEG activity to the past EEG activity and has been exploited in early studies on automated neonatal seizure detection. Gotman *et al*.,[25] detected rhythmic discharges by analyzing a frequency content of an epoch of EEG and comparing it to that of "background" epochs taken from 60secs before the epoch being investigated. Roessgen *et al*.,[26] defined the EEG spectrum as a summation of the spectra of background and seizure activity. Decisions were obtained using the ratio of the power in the seizure coefficients to the power in the background coefficients. Celka *et al*.,[6] proposed a seizure detector based on complexity analysis of the EEG. In this method the EEG was first pre-processed based on a model of background EEG to determine whether the signal structure corresponded to non-seizure or seizure activity. Recurrence quantification analysis was applied for similar purposed in adult EEG[27].

3.2. *The proposed adaptation*

A way to estimate how much the developed seizure detector algorithm is biased by non-seizure activity present in the ongoing EEG is presented in this section. The process of adaptation with the background model proposed in this work is shown in Fig. 4. The output of the SVM from channel j at time i, is converted to the probability $P(S|x_{ji})$. The probabilistic output is then compared to the threshold, θ, to detect non-seizure parts within the time interval, T, back from the current point i. The probabilistic output that corresponds to the non-seizure part is averaged to provide the estimated probabilistic level of background (non-seizure) EEG, δ_{ji}. It is then subtracted from the smoothed current probability of seizure, $\widehat{P}(S|x_{ji})$. Alternatively speaking, for a seizure event to survive the proposed adaptation, the probability of seizure is required to be greater than the threshold θ plus the current estimated level of the probability of the background δ_{ji}.

The scheme can be applied as shown in Fig. 4. However, for the chosen operating point (threshold, θ), the same decisions can effectively be obtained if applied only to those points that surpass the threshold, θ. It is worth noting that the proposed paradigm has no separate mechanism to identify 'non-seizure' activity. For any given user-defined threshold, a particular epoch is considered to be an epoch of non-seizure activity, if the probability output of the detector is below this threshold.

In fact, the proposed modification can be well formulated as a zero phase filter:

$$\widehat{P}(S \mid \mathbf{x}_{ji}) - \delta_{ji} = H(P)\,\widehat{P}(S \mid \mathbf{x}_{ji}) \quad (5)$$

where

$$H(P) = 1 - \frac{\delta_{ji}}{\widehat{P}(S \mid \mathbf{x}_{ji})} \quad (6)$$

It is now possible to draw a parallel between the proposed scheme and the spectral subtraction process proposed by Boll *et al*.,[28] for speech enhancement. During spectral subtraction, enhanced speech can be obtained by subtracting the estimated spectral components of the noise from the spectrum of the input noisy signal, assuming that the noise is stationary and is additive to the speech signal. In our work, rather than the spectral density domain, "spectral" subtraction is applied in the probability density domain. Background noise is first estimated on data that is considered to be non-seizure for the chosen threshold. This background estimate is then subtracted from the current probabilistic signal.

The designed $H(P)$ is a zero-phase filter, with its magnitude response in the range $0 \leq H(P) \leq 1$. The filter acts as a signal-to-noise (SNR) dependent attenuator. The attenuation at each pdf frequency bin increases with the decreasing SNR, and conversely decreases with the increasing SNR.

In an analogy to spectral subtraction, the scheme can be extended by introducing scaling factors α and β, with δ_{ji} and θ scaled by $\alpha \geq 0$ and $\beta \leq 1$, respectively, to control the harshness of the subtraction and the maximum level of the allowed background. In this work, the basic implementation was used with $\alpha=1$ and $\beta=1$. The parameter T which controls the number of classifier outputs taken into account depends on the stationarity of the noise. The T was estimated on the development data for each of the 18 runs in the LOO performance assessment. The median was around 10 minutes with insignificant variations in performance obtained in the 5-15 minute range. The $T = 10$ minutes was fixed for testing on the separate validation dataset.

It is important to notice that the modified system still works in the probabilistic domain, thus allowing for the control of the final decision by choosing different confidence levels, which makes the proposed system flexible for clinical needs.

4. Results and Discussion

4.1. *Leave-one-patient-out results*

ROC area—The per-patient ROC90 areas with and without background modelling are presented in Table 3. The absolute differences in per-patient performance in terms of ROC90 areas with their corresponding *p*-values are shown in Fig. 5. It can be seen from Fig. 5 that the separability of seizure and non-seizure probabilistic activity increases for patients which are heavily affected by respiration and sweating artifacts, patients 4, 6, and 18. The *p*-values of the two-tailed statistical significance test of the difference between the two ROC areas indicate that for patients 8, 9, 16, and 17 the proposed system modification has no statistically significant effects (p values>0.01). For patients 9 and 16, with the large number of seizures, some of which qualify as status epilepticus (with continuous seizures lasting more than 30m, see Table 1), there is no background estimated in the span of the last 10m from the current time, thus the current level of probability of seizure will not be subject to any subtraction and the performance will not be significantly altered. For the remaining patients, the corresponding *p*-values are close to 0 and here indicate statistically significant differences for these patients.

It is also worth noting that the number of testing datapoints (epochs) in each seizure patient is very large (as can be calculated from Table 1). According to (3), the standard error is inversely proportional to the square root of the product of the number of class datapoints. Moreover; the z statistic in (4) depends on the correlation coefficient r in the denominator which was 0.66 (median). Both factors; the large size of the dataset used and the correlation coefficient dictate that even the slightest differences in performance are statistically significant as shown in Fig. 5.

The overall positive effect of the proposed modification can be seen through the average ROC90 area across all seizure patients which has been increased from 77.9% to 82.4%. If the 20 non-seizure patients are taken into account when computing the specificity metric then the ROC90 area increases from 81.4% to 85.6%. The average ROC curve computed over the seizure patients is shown in Fig. 6. The overall ROC area has been increased from 95.7% to 96.5%.

From the results (as measured by the reported metrics) it can be determined that the trace of probabilistic derivatives is more discriminative than the trace of 'raw' probabilities. The proposed adaptation improves against respiration and sweating artifacts (patient 4, 6, and 18). At the same time, results on patients, which are known not to be affected by the respiration artifact also improved. To give an insight into the developed paradigm, we draw a parallel between the proposed method and spectral subtraction used in speech processing. In the latter, a speech detector is first used to detect speech; what is not detected as speech is then taken to represent noise. These noise components are then subtracted from the speech spectrum, which iteratively improves further speech processing. Essentially, the same process is used in this study. A window of recent EEG epochs which were not detected as seizure is used to estimate the 'noise' which is then subtracted from the seizure signal. One key difference in this study is that the adaptation is done in the probabilistic domain as compared to the spectral domain in speech. The adaptation is developed to get rid of estimated known additive noise introduced by stationary (in the probabilistic domain) respiration artifact. However, other patients' results improved from estimation and subtraction of unknown additive noise sources. Because our dataset is large, unedited continuous EEG, it is not feasible to fully annotate it in terms of all artifacts in order to determine exactly what types of artifacts are attenuated.

Seizure burden and FD/h—To estimate the burden the system would cause in a real clinical situation, the FD/h metric is calculated on all patients including 20 non-seizure patients.

The contribution of the background modelling approach to the clinically important metric is shown in Fig. 7. It can be seen that the number of false detection per hour is consistently lower when exploiting the designed paradigm for all operating points. In particular, the number of FD/h can be reduced from 0.42 to 0.18 while maintaining the correct detection of the seizure burden as high as 70%. Similarly, for a fixed FD/h rate of 0.25, the correct detection of the seizure burden can be increased from 65% to 73% with an additional 132 seizures detected.

An example of how the proposed adaptation of probability works is shown in Fig. 8 for illustration purposes. The output of the moving average filter is plotted for the same ~2 hours of EEG from patient 4 as used in Fig. 3. The EEG in the middle is corrupted with respiration artifact which results in an elevated seizure probability for the background as shown with a continuous white line. This initially results in a number of false detections, for the chosen threshold of 0.6 in this example which is indicated with a horizontal white line. Clearly, in this situation, any probabilistic spike should be considered relative to the past neighboring probabilities. After adaptation is applied, the modified threshold shown as a dotted white line cancels most false detections with the seizure activity still being correctly detected.

The adaptation is applied to every EEG channel separately. If seizure and respiration artifact happen to co-occur on the same channel, then it is likely that the seizure will be missed. The improved results indicate that co-occurring of seizure and respiration artifact is not a

common situation. If the respiration artifact and seizure occur in different channels at the same time, then the adaptation will cancel one and proceed with the other.

It is worth reemphasizing that unlike other studies which report performance increases obtained on datasets of several carefully selected minutes of EEG[6], the results in our study are obtained on the largest available dataset, which comprises 1479 hours of continuous unedited neonatal EEG, and thus these results are stable and significant. Moreover, the results reported in our work were not increased by averaging over training and testing data[8], nor by averaging over seizure and non-seizure patients[7], or by excluding worst performing patients[1] after their results have been obtained.

4.2. *Validation on the prospective dataset*

Along with the LOO performance assessment on the retrospective dataset of 38 neonates, the results of validation on the separate randomized truly unseen dataset of 51 neonates with and without the proposed adaptation are also presented in this study. The system was trained on 18 seizure patients from the LOO dataset. All the system parameters were fixed and a single test over the validation data was performed. It can be seen from Fig. 6 and Fig. 7 that the proposed adaptation similarly increases the performance of the algorithm. Quantitatively, the ROC90 and ROC were increased from 75.8% to 81% and from 93.4% to 96.1% respectively. This corresponds to a 41% relative improvement (96.1-93.4)/(100-93.4).

It can be seen that the results on the validation dataset are similar to the results reported using the LOO performance assessment in terms of both the epoch-based (ROC = 96.5% with LOO vs. ROC = 96.1% with validation dataset) and the event-based metrics (Sens@0.25 FD/h: 73% with LOO vs. 71% with validation dataset). In both cases, the validation performance curves fall within the confidence intervals of the performance reported with the LOO. This confirms a well-known fact that the LOO is the most unbiased prediction of the performance on unseen data. On the contrary, with a single split between training and testing[5,7,8] or without a separate testing dataset[10], the resulting performance significantly deviates from the performance obtained on a prospective validation dataset[1].

It is important to re-emphasize that no patients were excluded from the reported validation results here. Nor do we report median values to exclude the influence of worse performance on minority of patients. Both factors similarly lead to an artificial increase in the reported metric values.

There is one important difference between the validation results and the LOO results. A different neurophysiologist annotated the validation dataset. Alternatively speaking, the system was trained to match one ground truth standard and it was tested against another ground truth standard. To date there have not been any studies of inter-observer agreement in neonatal seizure detection on conventional EEG, however it is known that observers do not always agree[10,36]. In this context, the ability of the developed neonatal seizure detector to maintain reasonable accuracy and reliability under these adverse conditions has been tested and confirmed. Robust neonatal seizure detection becomes a concern as technology migrates from laboratory to field applications.

A separate clinical study is currently being prepared by our group to report the results of the ongoing validation campaign for regulatory approval on the multi-center evaluation dataset in complete detail and to further discuss the level of robustness of the developed system to various permutations in testing scenarios such as a change of recording environments, montage, recording personnel, clinical staff experience, and presence of various environment-specific non-biological artifacts.

Undertaking benchmarking evaluations has proven to be an extremely productive means for estimating and comparing algorithm performance and for verifying genuine technological advances in other areas of signal processing such as speech processing[30]. Official evaluations are an important vehicle for pushing the state-of-the-art forwards as it is only with standard experimental protocols and databases that it is possible to meaningfully compare different approaches.

5. Conclusions

A significant improvement in the performance of a patient independent neonatal seizure detector was achieved by the adaptive estimation of the probabilistic level of background activity. It was shown that the proposed background modelling is able to compensate for long lasting seizure-like artifacts, such as those due to respiration. It was shown that the inclusion of the background modelling into the probabilistic framework results in a significant increase in the seizure detection performance as measured by both epoch-based and event-based metrics. This background adaptation technique can be exploited in existing neonatal seizure detectors that produce continuous (probabilistic) output.

Acknowledgments

This work was supported in part by a grant from Science Foundation Ireland (10/IN.1/B3036). The authors would like to thank Dr. Nathan Stevenson for scoring the system on the validation dataset.

References

1. Cherian P, Deburchgraeve W, Swarte R, De Vos M, Govaert P, Van Huffel S, Visser G. Validation of a new automated neonatal seizure detection system: A clinician's perspective. Clin. Neurophysiol. 2011; 122:1490–1499. [PubMed: 21396883]
2. Temko A, Thomas E, Marnane W, Lightbody G, Boylan G. EEG-based neonatal seizure detection with support vector machines. Clin. Neurophysiol. 2011; 122:464–473. [PubMed: 20713314]
3. Rennie J, Boylan G. Treatment of neonatal seizures. Arch. Dis. Child. Fetal. Neonatal. Ed. 2007; 92:148–150.
4. Murray D, Boylan G, Ali I, Ryan C, Murphy B, Connolly S. Defining the gap between electrographic seizure burden, clinical expression and staff recognition of neonatal seizures. Arch. Dis. Child. Fetal. Neonatal. Ed. 2008; 93:187–191.
5. Mitra J, Glover J, Ktonas P, Kumar A, Mukherjee A, Karayiannis N, Frost J, Hrachovy R, Mizrahi E. A multistage system for the automated detection of epileptic seizures in neonatal electroencephalography. J. Clin. Neurophysiol. 2009; 26:1–9. [PubMed: 19151615]
6. Celka P, Colditz P. A computer-aided detection of EEG seizures in infants, a singular-spectrum approach and performance comparison. IEEE. Tran. Biomed. Eng. 2002; 49:455–462.
7. Navakatikyan M, Colditz P, Burke C, Inder T, Richmond J, Williams C. Seizure detection algorithm for neonates based on wave sequence analysis. Clin. Neurophysiol. 2006; 117:1190–1203. [PubMed: 16621690]
8. Aarabi A, Grebe R, Wallois F. A multistage knowledge-based system for EEG seizure detection in newborn infants. Clin. Neurophysiol. 2007; 118:2781–2797. [PubMed: 17905654]
9. Mitchell, M. Machine Learning. McGraw Hill; 1997.
10. De Vos M, Deburchgraeve W, Cherian P, Matic V, Swarte R, Govaert P, Visser G, Van Huffel S. Automated artifact removal as preprocessing refines neonatal seizure detection. Clin. Neurophysiol. 2011; 122:2345–2354. [PubMed: 21705269]
11. Chervin R, Burns J, Subotic N, Roussi C, Thelen B, Ruzicka D. Method for detection of respiratory cycle-related EEG changes in sleep-disordered breathing. Sleep. 2004; 27:110–115. [PubMed: 14998246]

12. Zhang A, Zheng C, Gu J. Removal of cardiac and respiratory artifacts from EEG recordings under increased intracranial pressure. Proc IEEE Machine Learning and Cybernetics. 2003; 4:2122–2126.
13. Park J, Jeong D, Park K. Automated detection and elimination of periodic ECG artifacts in EEG using the energy interval histogram method. IEEE Trans. Biomed. Eng. 2002; 49:1526–1533. [PubMed: 12549734]
14. Kitayama M, Otsubo H, Parvez S, Lodha A, Ying E, Parvez B, Ishii R, Mizuno-Matsumoto Y, Zoroofi R, Snead O. Wavelet analysis for neonatal electroencephalographic seizures. Pediatr. Neurol. 2003; 29:326–333. [PubMed: 14643396]
15. De Weerd A, Despland P, Plouin P. Neonatal EEG The International Federation of Clinical Neurophysiology. Electroencephalogr Clin Neurophysiol Suppl. 1999; 52:149–157. [PubMed: 10590984]
16. Temko A, Nadeu C, Marnane W, Boylan G, Lightbody G. EEG Signal Description with Spectral-Envelope-Based Speech Recognition Features for Detection of Neonatal Seizures. IEEE. T. Inf. Technol. Biomed. 2011; 15:839–847.
17. Temko A, Thomas E, Marnane W, Lightbody G, Boylan G. Performance assessment for EEG-based neonatal seizure detectors. Clin. Neurophysiol. 2011; 122:474–82. [PubMed: 20716492]
18. SVMlight - a collection of open-source software tools for learning and classification using SVM. http://svmlight.joachims.org/http://svmlight.joachims.org/
19. Platt J. Probabilistic outputs for SVM and comparison to Regularized likelihood methods Advances. Large Margin Classifiers. 1999
20. Vapnik, V. Estimation of Dependences Based on Empirical Data. Springer-Verlag; New York: 1982.
21. Mason S, Graham N. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Q. J. R. Meteorol. Soc. 2002; 128:2145–2166.
22. Hanley J, McNeil B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143:29–36. [PubMed: 7063747]
23. Hanley J, McNeil B. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983; 148:839–843. [PubMed: 6878708]
24. Walls-Esquivel E, Vecchierini M, Héberlé C, Wallois F. Electroencephalography (EEG) recording techniques and artefact detection in early premature babies. Clin. Neurophysiol. 2007; 37:299–309.
25. Gotman J, Flanagan D, Zhang J, Rosenblatt B. Automatic seizure detection in the newborn: Methods and initial evaluation. Electroenc. Clin. Neurophysiol. 1997; 103:356–362.
26. Roessgen M, Zoubir A, Boashash B. Seizure detection of newborn EEG using a model-based approach. IEEE Trans. Biomed. Eng. 1998; 45:243–246.
27. Acharya U, Sree V, Chattopadhyay S, Yu W, Alvin A. Application of recurrence quantification analysis for the automated identification of epileptic EEG signals. Int J Neural Syst. 2011; 21(3): 199–211. [PubMed: 21656923]
28. Boll S. Suppression of Acoustic Noise in Speech Using Spectral Subtraction. IEEE Trans. Acoust. Speech Signal Process. 1979; 27:113–120.
29. Vanhatalo S. Development of neonatal seizure detectors: An elusive target and stretching measuring tapes. Clin. Neurophysiol. 2011; 122:435–437. [PubMed: 20719559]
30. Rich Transcription Evaluation Project. National Institute of Standards and Technology (NIST). http://www.itl.nist.gov/iad/mig/tests/rt/http://www.itl.nist.gov/iad/mig/tests/rt/
31. Faust O, Acharya U, Min L, Sputh B. Automatic Identification of Epileptic and Background EEG Signals Using Frequency Domain Parameters. Int J Neural Syst. 2010; 20:159–176. [PubMed: 20411598]
32. Cabrerizo M, Ayala M, Goryawala M, Jayakar P, Adjouadi M. A New Parametric Feature Descriptor for the Classification of Epileptic and Control EEG Records in Pediatric Population. Int J Neural Syst. 2012; 22:1250001–16. [PubMed: 23627587]
33. Hsu WY. Continuous EEG signal analysis for asynchronous BCI application. Int J Neural Syst. 2011; 21:335–350. [PubMed: 21809479]

34. Acharya U, Sree V, Suri J. Automatic detection of epileptic EEG signals using higher order cumulant features. Int J Neural Syst. 2011; 21:403–414. [PubMed: 21956932]

35. Adeli, H.; Ghosh-Dastidar, S. Automated EEG-based Diagnosis of Neurological Disorders - Inventing the Future of Neurology. CRC Press, Taylor & Francis; Boca Raton, Florida: 2010.

36. Rennie J, Chorley G, Boylan G, 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; 89:37–40.

37. Shellhaas R, Soaita A, Clancy R. Sensitivity of amplitude-integrated electroencephalography for neonatal seizure detection. Pediatrics. 2007; 120:770–777. [PubMed: 17908764]

38. Adeli H, Ghosh-Dastidar S, Dadmehr N. A Wavelet-Chaos Methodology for Analysis of EEGs and EEG Sub-bands to detect Seizure and Epilepsy. IEEE. Tran. Biomed. Eng. 2007; 54:205–211.

39. Temko A, McEvoy R, Dwyer D, Faul S, Lightbody G, Marnane W. React: Real-time EEG analysis for seizure detection. AMA-IEEE Conf. Med. Tech. 2010

40. Adeli H, Zhou Z, Dadmehr N. Analysis of EEG Records in an Epileptic Patient Using Wavelet Transform. J Neurosci Meth. 2003; 123:69–87.

41. Faul S, Temko A, Marnane W. Age-independent seizure detection. EMBC. 2009

42. Temko A, Stevenson N, Marnane W, Boylan G, Lightbody G. Inclusion of temporal priors for automated neonatal EEG classification. J. Neural. Eng. 2012; 9

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

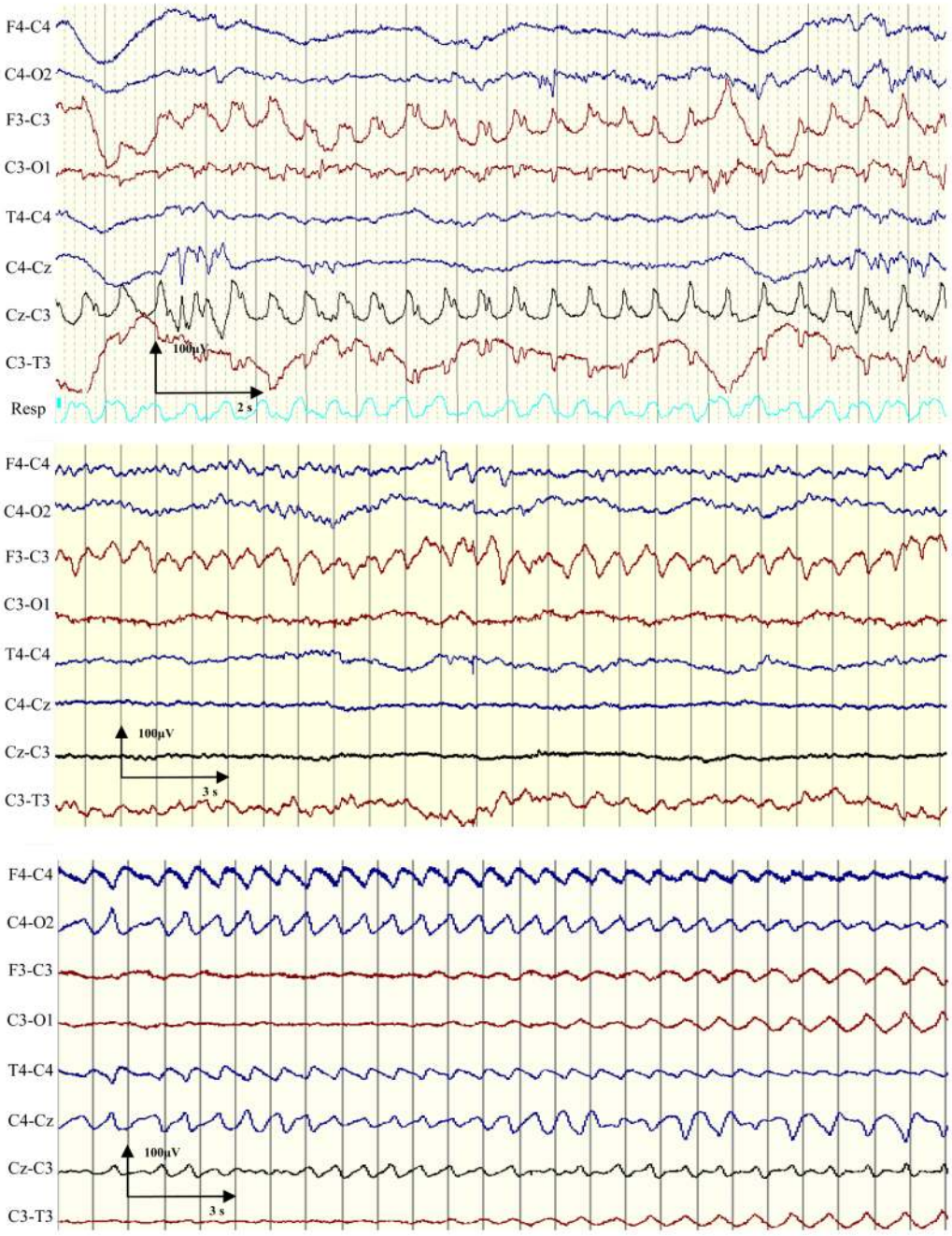

Fig. 1.

Neonatal EEG. Top graph shows channels F3-C3 and Cz-C3 corrupted by respiration artifact as can be seen from the respiration trace. Middle graph shows channels F3-C3 corrupted by respiration artifact which can be confused with seizure as respiration is not monitored. Bottom graph shows an example of an electrographic seizure in a baby with HIE. The seizure migrates from the right to the left cerebral hemisphere.

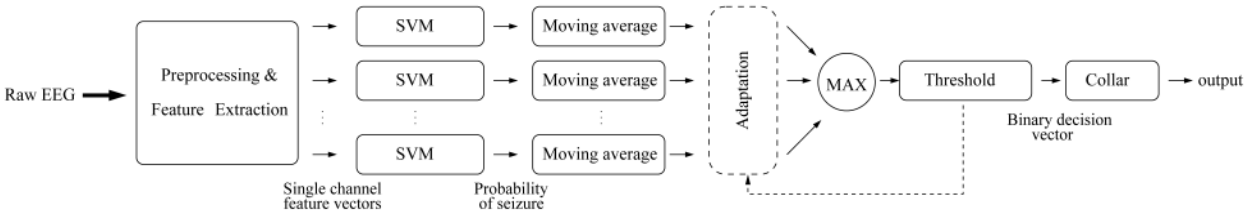

Fig. 2.
Neonatal seizure detection system diagram. Dashed line outlines the contribution of this work.

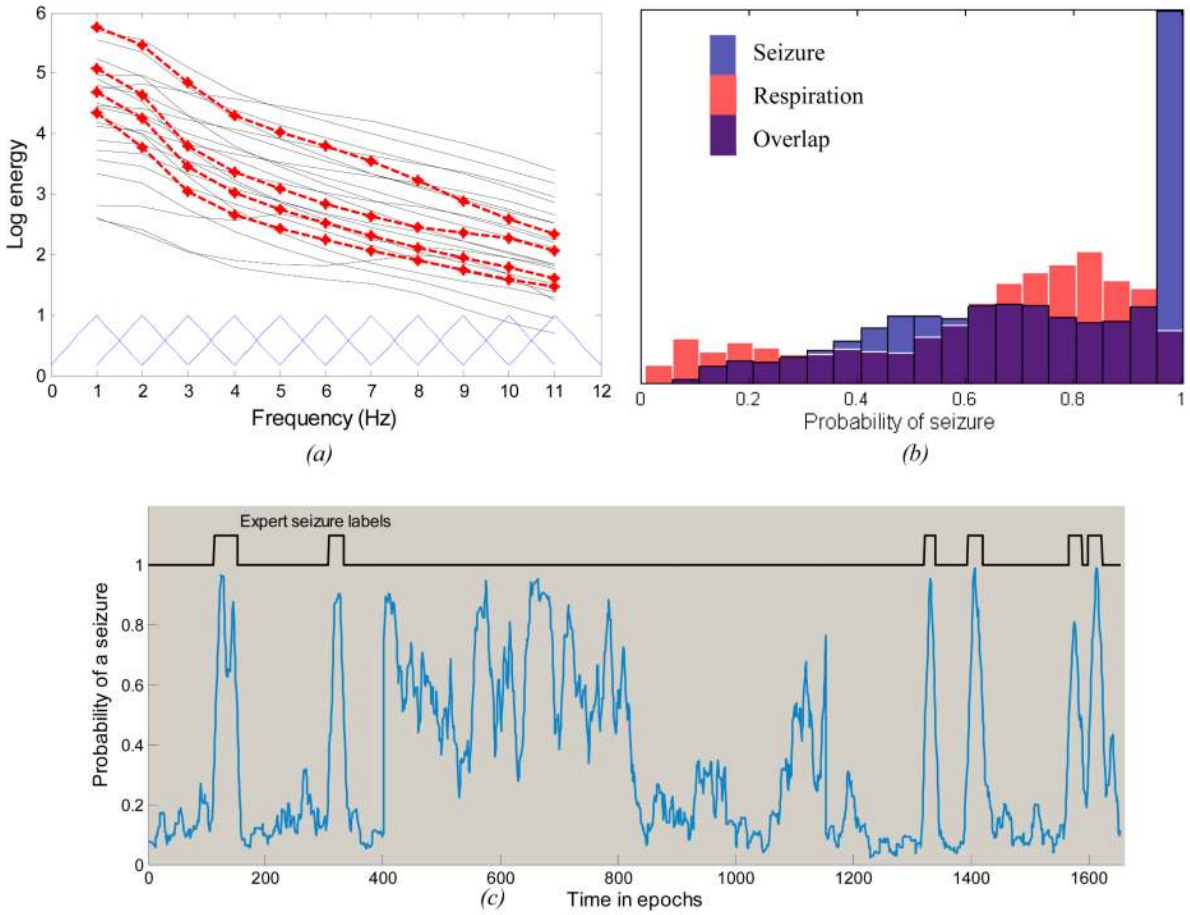

Fig. 3.

Example separation between respiration artifact and seizure. The (a) plot shows cluster centers of spectral envelopes; dashed line indicates those of respiration artifact. The (b) plot shows the histogram of the probabilistic output of the detector for the respiration class (red) and the seizure class (blue). The (c) graph shows an example of the probabilistic output with expert seizure labels superimposed on the top. The graph is best viewed in colour.

Europe PMC Funders Author Manuscripts

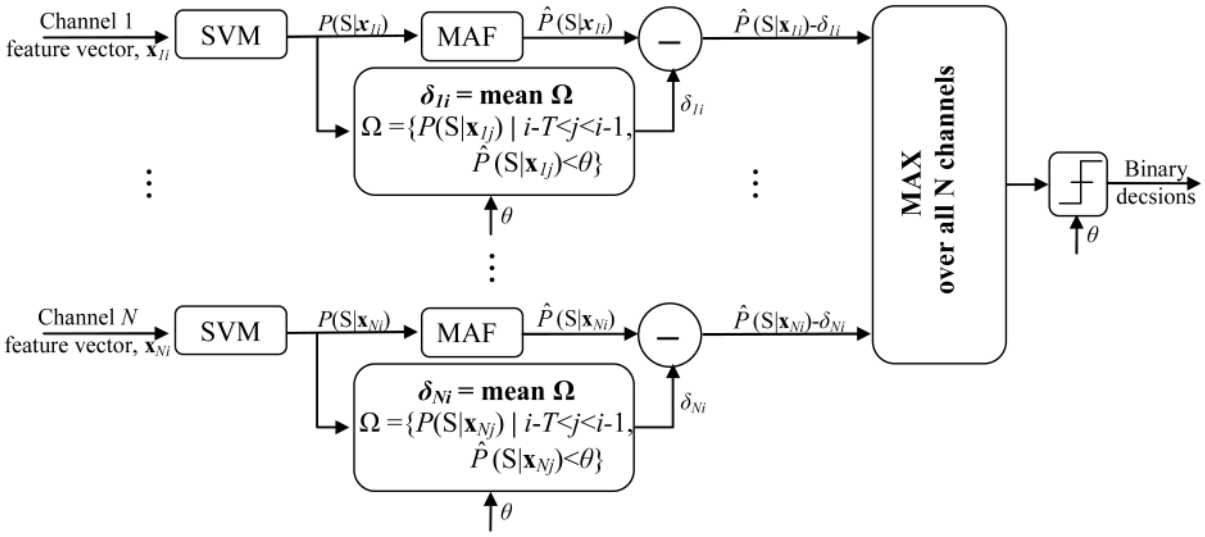

Fig. 4.
A flow chart of the proposed adaptation.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

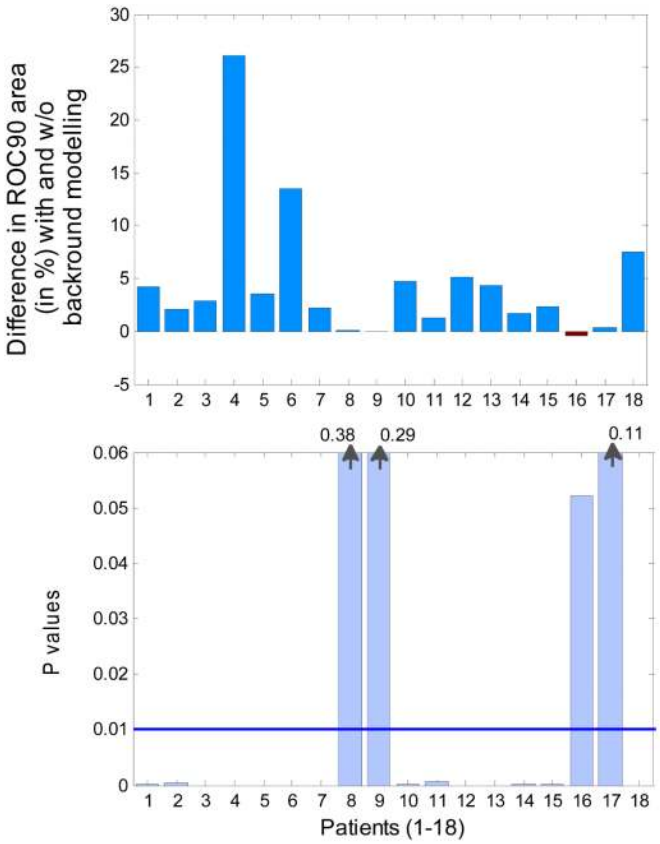

Fig. 5.

Absolute per-patient differences in ROC90 areas (top) and the corresponding *p*-values (bottom) for the neonatal seizure detector obtained with and without background modelling. Note that *p*-values for patients 8, 9 and 17 have been trimmed to fit the screen.

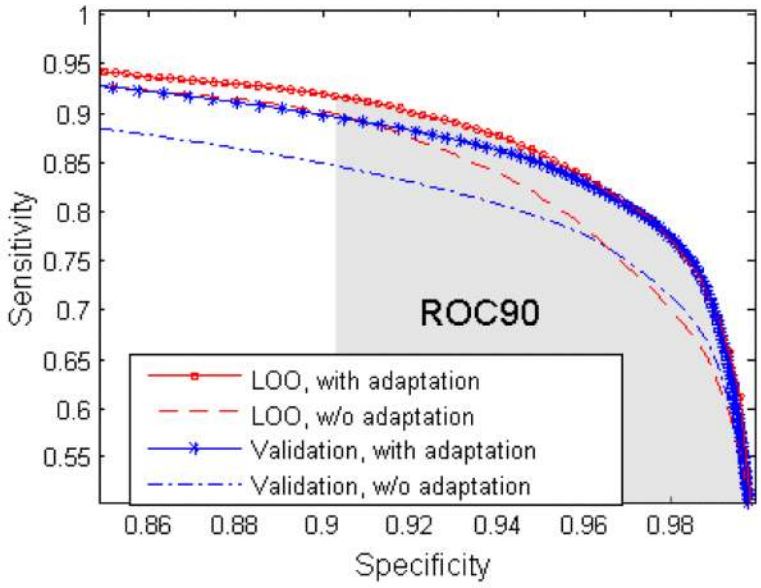

Fig. 6.
ROC area. ROC90 is indicated in grey.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

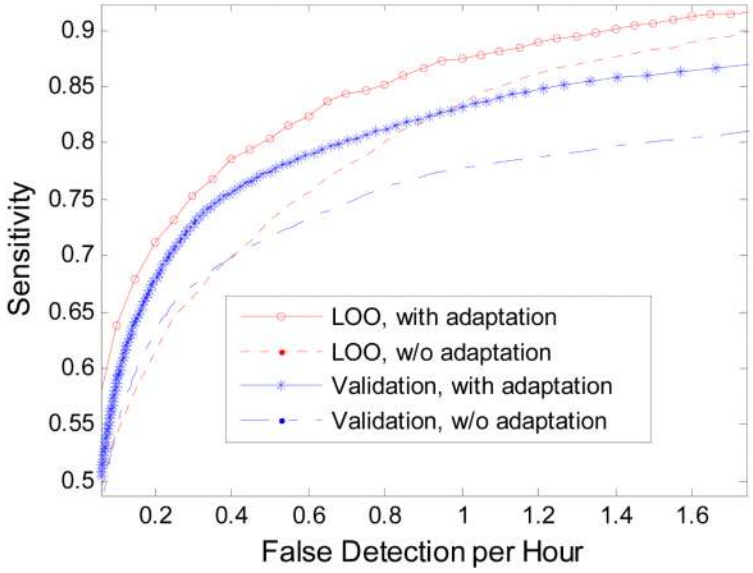

Fig. 7.
Sensitivity vs. number of false detection per hour.

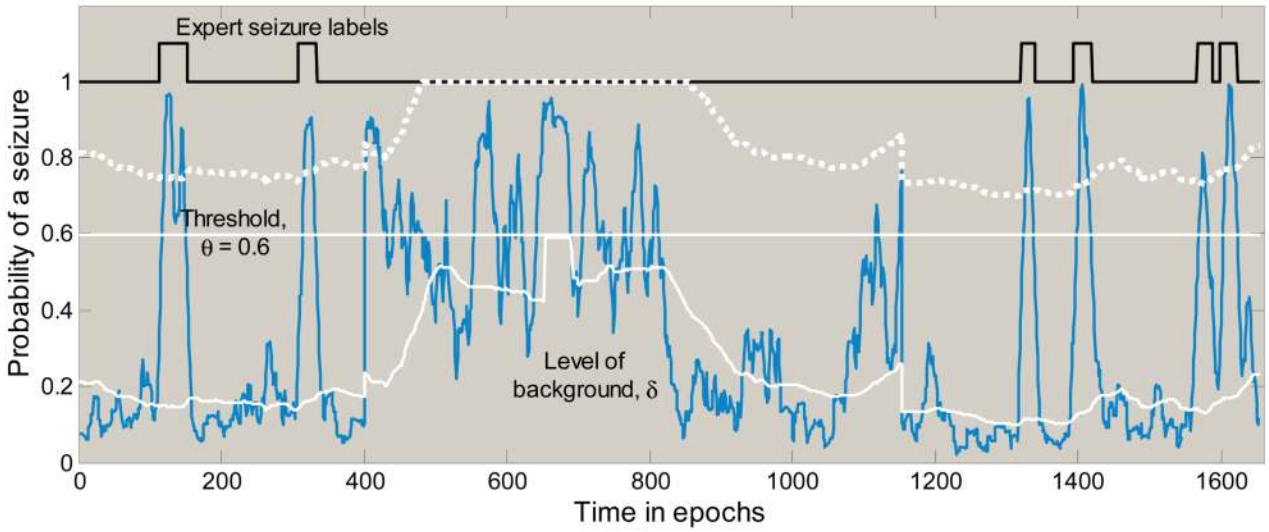

Fig. 8.

Example of the probabilistic output (in blue), $\hat{P}(S|x)$, for 1 hour 50 minutes of EEG from patient 4. The expert seizure labels are superimposed on the top. The threshold, θ, is set to 0.6 as indicated by the white horizontal line. The adaptively modelled level of background, δ ($\delta \leq \theta$), is indicated below the horizontal line, in white. The dynamic threshold for the current level of background, $\theta+\delta$, is indicated above the horizontal line, in white. The graph is best viewed in colour.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

EEG dataset of HIE newborns with seizures

Patient	Record length (h)	Seizure events	Seizure duration		
			Mean	Min	Max
1	29.7	17	1′30”	17”	3’54”
2	24.7	3	6’10”	55”	11’09”
3	29.9	209	1’50”	11”	10’43”
4	47.5	84	1’38”	32”	9’58”
5	47.2	62	6’37’	20”	34’10”
6	19.2	46	1’8”	15”	4’17”
7	60.8	99	1’32”	14”	10’20”
8	49.5	17	5’56”	29”	19’14”
9	67.7	201	4’59”	13”	37’06”
10	59.8	41	4’51”	13”	34’46”
11	21.8	43	2’27”	17”	7’36”
12	54.4	150	1’36”	15”	10’08”
13	51.7	60	3’26”	19”	16’56”
14	22.8	21	8’13”	22”	39’03”
15	59.7	121	1’31”	10”	7’08”
16	76.4	190	5’03”	26”	34’37”
17	30.7	21	5’31”	27”	23’16”
18	63.0	4	9’34”	7’19”	13’22”
Total	816.7	1389			

Table 2

Extracted features

Groups	Feature list
Frequency domain	• Total power (0-12Hz) • Peak frequency of spectrum • Spectral edge frequency (80%, 90%, 95%) • Power in 2Hz width sub-bands (0-2Hz, 1-3Hz, …10-12Hz) • Normalised power in sub-bands • Wavelet energy (the EEG is decomposed into 8 coefficients using the Daubechy 4 wavelet, the energy in the 5th coefficient corresponding to 1-2Hz is used as a feature)
Time domain	• Curve length • Number of maxima and minima • Root mean squared amplitude • Hjorth parameters • Zero crossings (raw epoch, Δ, ΔΔ) • Autoregressive modelling error (model order 1-9) • Skewness • Kurtosis • Nonlinear energy • Variance (Δ, ΔΔ)
Information theory:	• Shannon entropy • Singular value decomposition entropy • Fisher information • Spectral entropy

Europe PMC Funders Author Manuscripts

Table 3

LOO results on HIE newborns with seizures

Patient	ROC90 w/o adaptation	ROC90 with adaptation
1	78.9	**83.1**
2	91.4	**93.5**
3	78.7	**81.7**
4	53.6	**79.7**
5	66.3	**69.8**
6	59.5	**73.0**
7	89.3	**91.5**
8	80.8	**80.9**
9	**91.9**	91.8
10	57.0	**61.7**
11	87.4	**88.6**
12	71.3	**76.4**
13	84.7	**88.9**
14	83.8	**85.5**
15	85.0	**87.4**
16	**74.2**	73.8
17	90.9	**91.2**
18	77.8	**85.3**
Average	77.9	82.4