International Journal of Neural Systems, Vol. 28, No. 0 (2018) 1850011 (20 pages)

DOI: 10.1142/S0129065718500119

World Scientific
www.worldscientific.com

Neonatal Seizure Detection Using Deep Convolutional Neural Networks

Amir H. Ansari
Department of Electrical Engineering
KU Leuven, 3001 Leuven, Belgium
IMEC VZW, 3001 Leuven, Belgium

Perumpillichira J. Cherian
Department of Neurology, Erasmus University Medical Center
3015 CE Rotterdam, the Netherlands
Department of Medicine, McMaster University, Hamilton
ON, Canada L8S 4L8 Canada

Alexander Caicedo
Department of Electrical Engineering, KU Leuven,
3001 Leuven, Belgium
IMEC VZW, 3001 Leuven, Belgium

Gunnar Naulaers
Neonatal Intensive Care Unit
University Hospitals Leuven
Department of Development and Regeneration
KU Leuven, 3000 Leuven, Belgium

Maarten De Vos
Department of Engineering
University of Oxford, Oxford OX1 3PJ, UK

Sabine Van Huffel*
Department of Electrical Engineering
KU Leuven, 3001 Leuven, Belgium
IMEC VZW, 3001 Leuven, Belgium
sabine.vanhuffel@esat.kuleuven.be

Accepted 11 March 2018
Published Online 10 May 2018

Identifying a core set of features is one of the most important steps in the development of an automated seizure detector. In most of the published studies describing features and seizure classifiers, the features were hand-engineered, which may not be optimal. The main goal of the present paper is using deep convolutional neural networks (CNNs) and random forest to automatically optimize feature selection and classification. The input of the proposed classifier is raw multi-channel EEG and the output is the class label: seizure/nonseizure. By training this network, the required features are optimized, while fitting a nonlinear classifier on the features. After training the network with EEG recordings of 26 neonates, five end layers performing the classification were replaced with a random forest classifier in order to improve the performance. This resulted in a false alarm rate of 0.9 per hour and seizure detection rate

*Corresponding author.

of 77% using a test set of EEG recordings of 22 neonates that also included dubious seizures. The newly proposed CNN classifier outperformed three data-driven feature-based approaches and performed similar to a previously developed heuristic method.

Keywords: Deep neural networks; convolutional neural network; random forest; neonatal seizure detection.

1. Introduction

Neonatal seizures usually indicate serious neurological dysfunction, and could potentially worsen underlying brain injury.[1,2] The majority of neonatal seizures are acute symptomatic events, unlike the unprovoked epileptic seizures observed in older children and adults.[2,3] These seizures may have nonexistent or subtle clinical manifestations, which may resemble normal behavior, such as lip smacking, sucking, chewing, and blinking. This makes neonatal seizure detection very difficult and inaccurate if it solely relies upon clinical observation.[4–6] It has been shown that the most accurate method for their detection is visual interpretation of continuous multichannel EEG along with video by an expert clinical neurophysiologist.[1] However, such interpretation is extremely labor-intensive, time-consuming, and importantly, needs special expertise which is not available around the clock in many neonatal intensive care units (NICUs). A reliable and accurate automated neonatal seizure detector using multi-channel continuous EEG can be a very helpful supportive tool, particularly for the NICUs.

In the literature, there are several model-based methods for the automatic detection of neonatal seizures, usually developed based on heuristic if–then rules and some thresholds and parameters. Liu *et al.* computed the periodicity score using autocorrelation techniques and used three if–then rules and five thresholds to detect seizures.[7] Gotman *et al.* proposed a rhythmic discharge detector using three parallel methods in order to detect rhythmic discharges, multiple spikes, and very slow rhythmic seizures. This algorithm used 10 different thresholds in total.[8] Celka and Colditz used a singular spectrum analysis and compared the "optimum required model order" with a threshold to detect seizures.[9] Navakatikyan *et al.* applied a wavesequence analysis and used about nine thresholds to detect seizures.[10] Furthermore, a heuristic algorithm mimicking a human EEG reader was developed in our group, NeoGuard,[11,12] and was clinically validated.[13] In this method, spike-train-type and oscillatory-type seizures are detected by two parallel algorithms. The common feature of the aforementioned methods is that the thresholds and parameters were found empirically, usually by trial and error, and therefore they might not be optimized.

Other research groups focused on the development of data-driven methods for this task. Among them, the following algorithms have been considered: Hassanpour *et al.* used a singular value decomposition and "successive spike interval analysis" in order to extract, respectively, the low- and high-frequency features. Then, the features were fed into two separate artificial feed-forward neural networks, each of which has two hidden layers.[14] Greene *et al.* used 21 features, including frequency domain-, time domain-, and entropy-based features, extracted from 2 s epochs. These features then were used into a classifier based on linear discriminant analysis.[15] Thomas *et al.* applied a Gaussian mixture model (GMM) on 55 extracted features from 8 s epochs.[16] Temko and co-workers employed the same set of features with a support vector machines (SVMs) classifier, using a radial basis function (RBF) and a "Gaussian dynamic time warping" kernel.[17,18] A dictionary was created by Nagaraj *et al.* using an atomic decomposition technique applied on the training data. The complexity of the atoms was then measured and aggregated to define seizures.[19] Furthermore, Zwanenburg *et al.* extracted five features and used an SVM classifier to detect newborn lamb seizures.[20]

In addition, in some proposed methods, hybrid models combining data-driven methods, heuristic rules, and an empirically found set of parameters/thresholds have been used. Aarabi *et al.* used some if–then rules with predefined thresholds to remove artifacts, and then applied a multi-layer perceptron (MLP) with two hidden layers to detect seizures.[21] Furthermore, an MLP and a clustering method were used by Mitra *et al.* to detect and cluster seizures. Then, three rules with some predefined thresholds remove the artifacts and decrease the number of false detections.[22] Lastly, Ansari *et al.* developed a multi-stage classifier composed of a

heuristic method to detect potential seizures and a data-driven post-processor to remove artifacts. In the post-processor, different sets of features were introduced and extracted and an SVM was then used to classify the detected potential seizures.[23] In all the previously-mentioned data-driven approaches, the parameters of the classifiers were optimized by machine learning techniques. However, their performance was determined by the quality of the chosen features and finding appropriate features was a big challenge, which was typically performed by trial and error. This problem can be solved by using deep neural networks (DNNs).

In general, DNNs are referred to as artificial neural networks (ANNs) with several hidden layers.[24] Unlike the shallow (not deep) artificial neural networks used for seizure detection,[25–29] the deep networks do not need any hand-designed feature extraction unit. Different types of DNNs exist, e.g. convolutional neural network (CNN),[30] deep belief network,[31] stacked denoising auto-encoder,[32] long short-term memory (LSTM),[33] tensor deep stacking network,[34] etc. Among dozens of different DNNs, the convolutional neural network, CNN or ConvNet, has generated good results in image and speech processing applications.[30,35–39] Recently, these networks have also been found useful in EEG analysis: Cecotti and Graeser used a CNN embedded with a Fourier transform to classify steady-state visual evoked potential activities.[40] Furthermore, they also developed some CNN methods for detecting P300 responses in a brain–computer interface.[41] Mirowski *et al.* applied a previously developed CNN for document recognition, called LeNet, on seizure prediction data. They showed a significantly better prediction rate for the CNN method compared with an SVM-based and logistic regression approaches.[42] Acharya *et al.* proposed a new CNN method for automated seizure detection for adult patients. They designed a 13-layer network to process a single-channel EEG and achieved 95% sensitivity and 90% specificity.[43]

Recently, O'Shea *et al.* proposed a single-channel CNN for neonatal seizure detection.[44] In this method, the network uses 8 s of a single-channel EEG signal as input to the CNN. Then, a post-processor is applied on the outputs of the CNN. However, we consider that 8 s are not enough for extracting evolutionary features of EEG, which have been shown to be important EEG characteristics for discrimination of brief-lasting seizures (<30 s) from short artifacts.[23]

This paper introduces a seizure detection algorithm using CNN, specifically for neonatal seizures, which takes a segment of raw multi-EEG data and then labels it as seizure or nonseizure. Unlike most previously proposed methods in this field, this method does not need preprocessing of the EEG data, hand-engineered feature extraction procedures, or complex post-processing to aggregate epochs or channels. It automatically extracts the best-required features and classifies each segment of raw multi-channel EEG based on those features. Once the CNN is trained, it can be merged with other classifiers, such as LSTM,[45,46] random forest (RF),[47] SVM,[48,49] etc., to improve its performance.

In this paper, the proposed CNN is merged with an RF to detect neonatal seizures from 90 s multi-channel EEG segments. This method is compared with our previously developed heuristic approach, as well as with three feature-based data-driven algorithms (Algorithms 1–3).

The main objective of this paper is to introduce a CNN-based algorithm and compare it with hand-designed, feature-based, and heuristic methods with no complex pre/post-processing steps. Dozens of pre/post-processing algorithms exist for improving the performance of neonatal seizure detectors proposed in the literature. For instance, De Vos *et al.* used blind source separation techniques for removing artifacts as a preprocessor.[12] Temko *et al.* applied adaptive background modeling to adaptively change the latent variable with respect to the background activity as a post-processor.[50] A data-driven post-processor was proposed by Ansari *et al.* to find evolutionary patterns of seizures to distinguish between the real seizures and polygraphic signals-related artifacts (e.g. ECG artifacts).[23] They also used an adaptive learning technique to apply the experts' feedback to tune the latent variable.[51] These algorithms can also be applied on the outputs of the methods considered in this paper in order to improve their performance. However, this is outside the scope of this paper.

The paper is organized as follows: the database and the used methods including a heuristic, three feature-based, and the proposed CNN methods are explained in Sec. 2. The results of the methods are

reported and compared in Sec. 3. Discussion is given in Sec. 4 and conclusions are drawn in Sec. 5.

2. Materials and Methods

2.1. *Database*

EEG recordings from 48 newborn babies were used to train and test the algorithms. These recordings were obtained at the NICUs of Sophia Children's Hospital (part of the Erasmus University Medical Center, Rotterdam, the Netherlands) between 2003 and 2012. All the subjects were termed neonates (with gestational age $\geq$ 36) and admitted to the NICUs with presumed postasphyxial hypoxic ischemic encephalopathy (HIE), all underwent continuous EEG monitoring. Inclusion criteria were either a 5 min Apgar score below six or an umbilical artery pH $<$ 7.10, and clinical encephalopathy according to Sarnat score.[3,52–54] Five hours of recordings, on average, were used for each neonate, in which at least one seizure was observed. The seizure periods were scored by an expert clinical neurophysiologist and annotated as seizure when a clear change in the background EEG activity lasting for at least 10 s was observed with evolution in amplitude and/or frequency.[13] For each annotated seizure, the onset and offset were indicated by the expert. The dubious seizures were not removed from the database and no preselection has been performed based on the presence of artifacts or quality of signal.[13,55] Newborns with brain or heart malformation were excluded for this study. All recordings were fully anonymized. The Erasmus MC Medical Ethics Committee approved a study (2003–2007) to assess the utility of continuous EEG monitoring in neonates with postasphyxial HIE. Use of anonymized EEG data from this study, for analysis and research, was subsequently approved.

From the 48 EEG recordings, 39 include "Fp1–2," "F7–8," "T3–4," "T5–6," "O1–2," "F3–4," "C3–4," "P3–4," and "Cz" electrodes (17 electrodes) [Fig. 1(a)]. In seven recordings, "F3–4" and "P3–4" were not available (13 available electrodes) [Fig. 1(b)]. In the two remaining recordings, "F7–8," "T5–6," "F3–4," and "P3–4" were not recorded (nine available electrodes) [Fig. 1(c)]. In order to obtain bipolar channels, a full and two restricted montage maps were used [Fig. 1(a)–1(c)].[56] As a result, 20, 12, and 12 bipolar channels were derived, respectively, using the aforementioned 17, 13, and 9 electrodes. In addition to EEG signals, all recordings include polygraphic signals, such as electrocardiogram (ECG) and electro-oculogram (EOG), which were not used in this study. The initial sampling frequency for the measurements was 256 Hz.

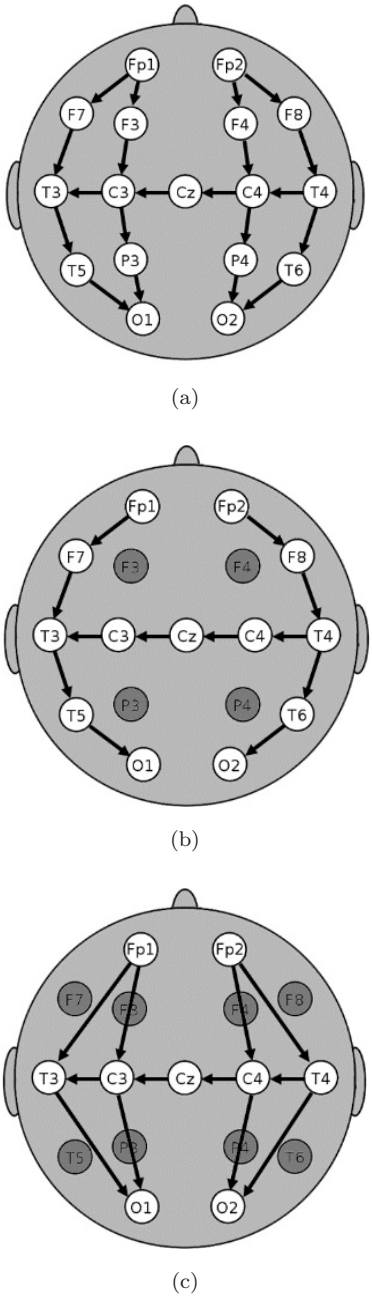


Fig. 1. Neonatal EEG montages. (a) Full 10–20 system of electrode placement using 17 electrodes, (b) a restricted 10–20 system using 13 electrodes, and (c) a restricted 10–20 system using nine electrodes.

2.2. *Segmenting*

In order to train the different classifiers, 4344 segments (50% seizure and 50% nonseizure) have been selected manually from 26 neonates. Note that some segments can correspond to different parts of the same seizure or even can overlap with each other. Each of these segments includes 90 s of EEG data from all available bipolar channels. In order to have a sufficient number of training data, the data was segmented using consecutively overlapping segments with overlaps of 2, 4, and 6 s to the left and the right, obtaining a total of more than 30,000 segments to be used for training. The training data were split into training (75%) and validation (25%) sets to stop the backpropagation algorithm. For the test dataset, the whole recordings of the 22 remaining neonates were split into 90 s segments with 60 s overlap. In this method, a window length of 90 s was chosen, since this length is considered long enough to extract the dynamics and evolutionary characteristics of brief-lasting seizures ($<$ 1 min), which are reported as the most difficult seizures for automatic detection,[18,19,23,51,57] and is not too wide to avoid a too long delay between the onset of seizures and the alarm (the maximum delay equals the window length, 90 s). None of the testing neonates or segments has been used in the training process. All data-driven methods, which will be explained in this section, used these training and test datasets. In contrast, the heuristic algorithm did not need training data. However, part of this training dataset was used by Deburchgraeve *et al.* to tune the parameters and thresholds.[11] To guarantee full independence with the test set, the EEG data from neonates previously used for developing the heuristic method were excluded in any test performed with the classifiers considered in this paper.

2.3. *Heuristic method*

In this paper, we used a previously developed heuristic model that mimics a human EEG reader to compare with the proposed CNN method. This algorithm was developed in our group by Deburchgraeve *et al.*[11] and its schematic overview is displayed in Fig. 2. The comprehensive description of its last version, which is used here, is available in Appendix A of the original paper.[12] Briefly, this algorithm uses two separate procedures for detecting

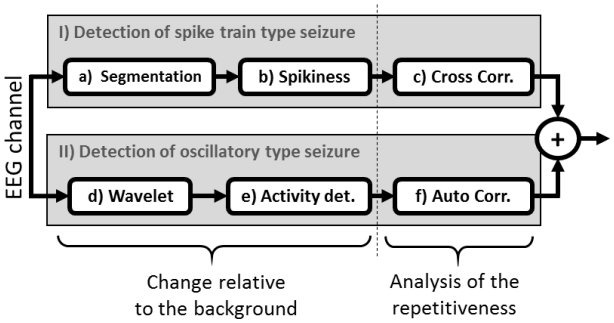


Fig. 2. Schematic overview of the heuristic method. The upper and lower lines show the spike-train-type and oscillatory-type seizure detectors, respectively.

seizures: (I) a spike-train seizure detector and (II) an oscillatory seizure detector. In the spike-train detection, first, a nonlinear energy operator (NLEO), using the Teager–Kaiser operator, is applied on one channel of EEG data and the output is normalized and smoothed with a moving average (MA) filter with a window size of 120 ms. Then, an adaptive threshold is applied and the potential spikes which have a smoothed nonlinear energy greater than a specified threshold are selected [see (a) in Fig. 2]. Next, the selected segments with a duration of more than 60 ms and isolated from the background activity are detected as epileptic spikes [see (b) in Fig. 2]. Finally, when at least six sequential spikes have the overall cross-correlation higher than 0.8, they are considered as a spike-train-type seizure [see (c) in Fig. 2]. In the second detector, the δ (0.5–4 Hz) and θ (4–8 Hz) frequency bands are extracted from one channel of EEG data using a discrete wavelet transform [see (d) in Fig. 2]. Then, the potential epileptic activities are defined when 3 s of filtered signal has significantly higher energy compared to its previous 30 s [see (e) in Fig. 2]. Next, autocorrelation analysis and two thresholds are applied on the potential activities to detect subsequently the oscillatory seizures [see (f) in Fig. 2]. These two procedures are applied on all channels of EEG individually. If there is a seizure in at least one channel, that segment is marked as seizure ("OR" operator).

2.4. *Feature-based approaches*

In order to compare the classic feature-based approaches to the proposed CNN-based method, which needs no predefined features, three feature-based algorithms are proposed (Algorithms 1–3). All

Table 1. Extracted features used in Algorithms 1–3.

Type	Feature name (number)	Short description
Frequency Domain	Total power (1)	The total power of estimated power spectral density (PSD) in the range of 1–20 Hz
	Peak frequency (1)	The peak frequency of the PSD
	Spectral edge frequencies (3)	The frequencies below which 80%, 90%, and 95% of the total spectral powers are kept
	Spectral power (11)	The spectral power of 11 specific bands including (0–2 Hz, 1–3 Hz, . . . , 10–12 Hz)
	Normalized power (11)	The normalized spectral power of the same 11 bands
	Wavelet energy (1)	The energy of the wavelet coefficients in the 5th level of decomposition using Daubechies-4 (corresponding to 1–2 Hz)
Time Domain	Line (curve) length (1)	The sum of the absolute values of the differences in amplitudes of consecutive samples
	Root-mean-squared amplitude (1)	The root-mean-squared value of the epoch
	Hjorth parameters (3)	The Hjorth activity, mobility, and complexity metrics
	Zero crossing (3)	The number of zero crossings of the EEG, as well as its first and second derivatives
	Variances (2)	The variances of the first and second derivatives of the epoch
	Skewness and kurtosis (2)	The skewness and kurtosis of the epoch
	Nonlinear energy (1)	The averaged nonlinear energy using Teager–Kaiser energy operator
	Number of maxima and minima (1)	The total number of local minima and maxima in the epoch
	Autoregressive modeling error (9)	The error of autoregressive modeling with order 1–9
Info. Theory	Spectral entropy (1)	The normalized spectral entropy using PSD
	Shannon entropy (1)	The Shannon entropy using histogram of the data distribution
	SVD entropy (1)	The entropy of normalized singular values of the EEG epoch
	Fisher information (1)	The Fisher information of the EEG epoch

the algorithms used the same feature set including 28 features from the frequency domain, 23 from the time domain, and four from information theory (in total 55 features), which are listed in Table 1. These features have been used in different methods for neonatal seizure detection.[8,15–18,21,50,57–59] More information about the computation of these features can be found in the reference literature.[8,15,21,59] The classifier used in these algorithms is a bagged random forest. However, the splitting of EEG or the aggregating of the channels is different depending on the algorithms.

Algorithm 1. Fifty-five features from each channel (each 90 s) are calculated and concatenated to make a vector with 1100 features in total (= 55 features ×20 channels). Then, these features are fed into the classifier. The probabilistic output of the classifier is compared with a threshold to define the label. Since the number of channels should be constant to result in a fixed input size, a zero vector is used for the unavailable bipolar channels.

Algorithm 2. In this algorithm, the features are extracted and classified for each channel separately. Therefore, the input of the classifier is 55 features extracted from each individual channel. Then, the probabilistic output of the classifier for each channel is compared with a threshold. If at least the output of one channel is greater than the threshold, the whole segment is considered as seizure ("OR" operator on channels).

Algorithm 3. In this algorithm, first, the EEG data of each channel was split into 8 s epochs with

50% overlap. Second, 55 features of the epochs are extracted and the classifier is applied. Third, the probabilistic outputs of all epochs in each channel are smoothed with a moving average filter ($N = 15$) and compared with a threshold. If at least one epoch of a channel contains a seizure, the whole segment is considered as seizure ("OR" operator on epochs and channels). The general idea of this method was derived from the method proposed by Temko *et al.*[18] However, there are two differences between them: (1) instead of using an SVM classifier, an RF is used in order to have a fair comparison with the proposed method which uses an RF. For our dataset, the RF method results in a higher performance than the SVM classifier as is reported in Sec. 3. (2) The collar method used by Temko *et al.* for correcting the onset and offset of detections is not used here. As mentioned before, the previously proposed pre/post-processing methods (e.g. collar) can be similarly applied on the CNN or feature-based approaches to improve the performance.

In Algorithms 1–3, the mentioned threshold is varied from 0 to 1 in order to construct the receiver operating characteristic (ROC) curve. Since in Algorithms 2 and 3 the classifier is respectively applied on channels and epochs of a segment separately, in order to improve the training, the exact moment and the channels representing seizures were premarked in each segment of the training dataset by the method developer.

2.5. *Proposed CNN–RF method*

We propose the use of a CNN for the automatic detection of epileptic seizures. As mentioned before, the main advantage of the proposed method is that there is no need to select any features manually. In other words, the classifier takes the raw multi-channel EEG data and automatically optimizes the features and classifier at the same time. In order to improve the classification performance, when the CNN was trained, the classifying end layers were removed and replaced by an RF. In this method, an RF classifier was selected since it performs better than other classifiers. In order to test this, a bootstrap test was applied on the training data. First, the training data was split into 75% training and 25% validating subsets. Then, four classifiers including LDA, two SVMs (with linear and RBF kernels), and an RF were trained. A 10-fold cross-validation and grid search was used to optimize the hyper-parameters of the SVM–RBF and RF methods. Next, the area under the curve (AUC) of each classifier on the validation set was calculated. The splitting has been repeated 100 times. As presented in Sec. 3, the RF has a significantly better performance than others. As a result, in the final model, the CNN is considered as an automatic feature extractor and the RF is the classifier (Fig. 3). The following sub-subsections describe this approach in detail.

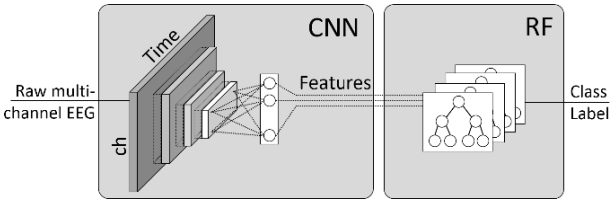


Fig. 3. Schematic overview of the proposed CNN–RF method.

2.5.1. *Overview of CNN*

CNNs are classified as a special type of feed-forward artificial neural networks. In general, a CNN consists of multiple stacked layers of three different types: convolutional layer (Conv), nonlinear layer, and pooling layer. Note that the input of each layer is a three-dimensional volume.

Conv layer. This layer, which is the main block of CNN, is composed of a bank of finite impulse response (FIR) filters (also called kernels) that operate on the input as follows:

$$O(i,j,k) = \sum_{p=1}^{P} \sum_{n=1}^{N} \sum_{m=1}^{M} f_k(m,n,p) I(i-m, j-n, p), \tag{1}$$

where $I(i,j,p)$ is the input of the Conv layer, where (i,j,p) represents the dimensionality of the input data and $f_k(m,n,p)$ are the coefficients of the kth filter which consists of $M \times N \times P$ coefficients, where M and N represent the size of the filters and P represents the number of filters in the previous layer. $O(i,j,k)$ is the output of Conv layer, resulted from the convolution operator of the filter f_k and the input I through the first and second modes. The filter coefficients, f_k, are the only unknown parameters of the CNN which should be found in the training process

Table 2. Layers of the designed network before pruning.

		Layer info	Output size
Input:			(20, 2700, 1)
Feature Extraction	1	Conv$(1,5)\times 5$	(20, 2696, 5)
	2	MPool$(1,3), s:2$	(20, 1347, 5)
	3	ReLU	(20, 1347, 5)
	4	Conv$(1,5)\times 8$	(20, 1343, 8)
	5	MPool$(1,3), s:2$	(20, 671, 8)
	6	ReLU	(20, 671, 8)
	7	Conv$(1,5)\times 10$	(20, 667, 10)
	8	MPool$(1,3), s:2$	(20, 333, 10)
	9	ReLU	(20, 333, 10)
	10	Conv$(1,5)\times 15$	(20, 329, 15)
	11	MPool$(1,3), s:2$	(20, 164, 15)
	12	ReLU	(20, 164, 15)
	13	Conv$(1,20)\times 20$	(20, 145, 20)
	14	MPool$(1,10), s:5$	(20, 28, 20)
	15	ReLU	(20, 28, 20)
	16	MPool$(1,5), s:3$	(20, 8, 20)
	17	APool$(1,8), s:1$	(20, 1, 20)
	18	MPool$(20,1), s:1$	(1, 1, 20)
Classifier	19	Conv$(1,1)\times 5$	(1, 1, 5)
	20	Sigmoid	(1, 1, 5)
	21	Conv$(1,1)\times 2$	(1, 1, 2)
	22	Sigmoid	(1, 1, 2)
	23	Loss	(1, 1, 1)
		Total number of parameters: 7600	

Notes: Conv: Convolutional layer, the information is given in the following format (dimension in channel, number of coefficients in time) ×number of filters.
MPool: Pooling by maximum operator, the information is given in the following format (dimensions in channel, number of coefficients in time), s: stride.
*A*Pool: Pooling by average operator.
ReLU: Rectified linear unit.
Loss: Loss function.

by a backpropagation method. However, the size of filters (M, N), known as receptive field, as well as the number of filters of each Conv layer should be predefined in the design process (hyper-parameters). The output size of the Conv layers in the first and second modes resulting from the convolution operator equals the size of the input subtracted by the length of the filter plus one. The output size of the third mode is equal to the number of filters in that layer. For instance, in the proposed method, see Table 2, the first Conv layer is composed of five filters each one of them of size 1×5. The size of the output of this layer in the second mode is $2696 (= 2700 - 5 + 1)$.

Comparing the FIR filters of the Conv layer with common neurons in ANN shows that each filter is like a layer of simple linear neurons with two important characteristics: (1) the weights of all neurons located in the layer are shared between the neurons and (2) neurons only connect to a limited number of inputs with overlap. Applying these two characteristics on a layer of simple neurons converts the layer to the mentioned convolutional FIR filter.

Pooling layer. The main aim of pooling layers is reducing the number of outputs of the Conv layers by a nonlinear subsampling function in local regions. In practice, taking the maximum and averaging are the

two most common operations being used in pooling layers. The stride of pooling should be predefined as a hyper-parameter. The output volume size of this layer in each mode equals

$$\left\lfloor \frac{\mathrm{S}_{\mathrm{input}} - S_{\mathrm{filter}}}{\mathrm{stride}} \right\rfloor + 1, \tag{2}$$

where S_x represents the size of x and $\lfloor\ \rfloor$ is the floor function. For instance, in the proposed method, see Table 2, the pooling in the eighth layer is a 1×3 max pooling with stride 2 which decreases the size of the second mode from 667 to 333. Note that pooling layers have no trainable parameters.

Nonlinear layer. This is a nonlinear unit that increases the nonlinearity and power of the network. The most commonly used function in CNNs is rectified linear unit (ReLU) which is defined as follows:

$$O(x) = \max(0, x), \tag{3}$$

where x is the input value and $O(x)$ is the output. In other words, this function is a half-wave rectifier which replaces the negative values of the Conv layer output with zero. It has no effect on the size of data. In addition, in the very last layers of CNN, which are performing the classification task, where the first and second modes are completely aggregated by pooling, the Sigmoid unit is also suitable which is computed by

$$O(x) = \frac{1}{1 + e^{-x}}. \tag{4}$$

For instance, in the proposed method, the ReLU and Sigmoid units were used in the 12th and 22nd layers, respectively.

2.5.2. *Structure of the proposed CNN*

The input of the proposed CNN is first filtered between 0.5 Hz and 15 Hz. In order to decrease the complexity of the CNN network, the EEG is downsampled to 30 Hz. As a result, each segment of EEG, which is 20 channels by 90 s, is converted to an image with a size of 20×2700 where the first and second modes correspond to channel and time. For the neonates having a fewer number of channels, zero vectors are used to make a homogeneous input image with the same size. These images are the inputs of the proposed CNN.

The structure of CNN is formed by 23 layers, which are listed in Table 2. In this table, the dimensions of the filters as well as the number of Conv filters, stride of pooling layers, and the output size of each layer are shown. The first 15 layers are composed of five blocks of (Conv + max pool + ReLU) in order to extract the features related to seizure patterns. The beginning blocks extract local abstractions, like the slope of lines, whereas the deeper blocks extract more global ones, such as the spike and oscillation patterns. In layers 16 and 17, a maximum and an averaging pooling layer aggregate all time samples. It means that each output of the 17th layer represents the abstraction of the whole 90 s of the corresponding channel. Due to the fact that in our database, the recording of different neonates included a different number of electrodes and subsequently different bipolar channels all Conv and pooling layers in these first 17 layers are $1 \times N$ operators. It means that they filter and aggregate only the time information, and have no effect on channel (spatial) information. In other words, in these layers, the process is performed on each channel separately. Next, in layer 18, a max pooling is performed on all 20 channels in order to aggregate the features of different channels. The used maximum operator, which can be considered as an "OR" operator in fuzzy logic, for the channel aggregation is supported by the fact that in a clinical neurophysiologists' point of view, if one channel of EEG represents a seizure pattern, the whole segment is marked as seizure. As a result, the outputs of the 18th layer are 20 numbers (features) representing the characteristics of all channels and the whole 90 s. Then, the remaining five fully connected layers including two hidden Conv layers, two Sigmoid nonlinear units, and finally a loss function for computing the classification error are performing the classification task. This structure as well as the hyper-parameters were chosen by trial and error. When the network is trained, these end five layers are removed and replaced with an RF. Figure 4 schematically shows the designed layers, features, and the fully connected classifier.

The CNN implementation was performed by the MATLAB toolbox MatConvNet.[60] For training the CNN, the weights of the Conv layers were initiated with normally distributed random numbers generated by $N(0, \sigma^2)$, where the standard deviation, σ, equals 0.2 and 0.1 for layers 1–18 and 19–23, respectively. All bias weights of the Conv layers were initiated by zero. The learning rates were varying from 0.3 to 0.003 with respect to the layer

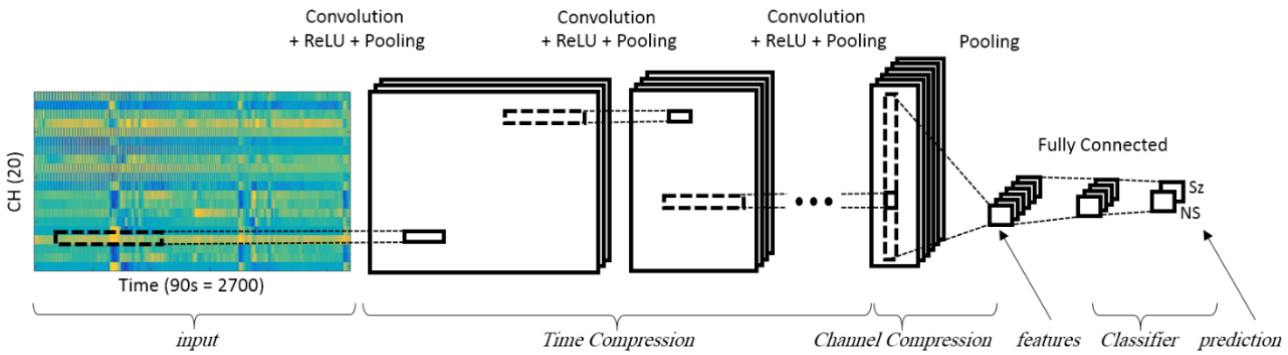


Fig. 4. Schematic structure of the proposed CNN method.

and epoch numbers. The learning batch size was 20 segments.

2.5.3. *Pruning and tuning*

As explained previously, an ReLU replaces negative outputs of a Conv layer with zero. Thus, if the output of an ReLU is always zero, for all seizure and nonseizure segments of the training dataset, it means that the corresponding filter of the Conv layer located before that ReLU is always producing negative outputs and, therefore, has no influence on the final output of the network. In order to rank the effectiveness of the filters and prune them with respect to the aforementioned fact in the presence of outliers, the 99% percentile, p_{99}, of the outputs of each filter is calculated through the training dataset. If the p_{99} is negative, that filter, as well as all corresponding parameters of the Conv layer of the next layer, is removed. For positive but small values of p_{99}, the procedure is continued till the validation error increases. The layers and parameters of the pruned CNN are listed in Appendix A (see Table A.1). The total number of parameters of the CNN reduced by 58% after pruning, which increases the generalization power of the network as well as the training and recall speed. When the selected filters and parameters are removed, the network was retuned by the training data for a few extra epochs.

2.5.4. *Using random forest*

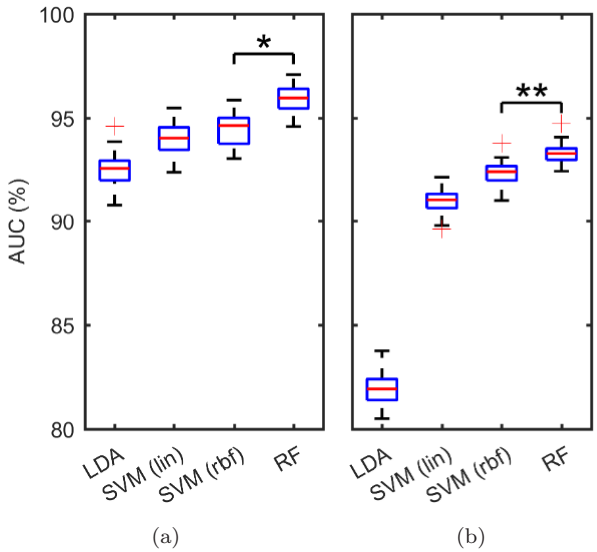
When the CNN is trained, the last five classifying layers were replaced with an RF classifier. Therefore, the first 18 layers of the CNN act as an automatic feature extractor. The RF is composed of 100 bagged decision trees. In order to train each decision tree, first, $\sqrt{N}$ features, where N is the total number of features extracted by the CNN, are randomly chosen. Second, a new set of Q data points is created from the Q available training segments using random selection with replacement; this procedure is normally called bagging. Then, a decision tree was trained using the Q segments and $\sqrt{N}$ features based on "classification and regression tree" analysis, namely CART.[61] Briefly, first, all possible binary splits of each feature are performed and the Gini's diversity index (GDI) of the tree after each split is calculated. Second, the split that has maximized the GDI is selected and the two consequence child leaves are formed. The procedure recursively repeats for each leaf until one of the following stopping conditions is reached: (1) when the tree depth equals a predefined maximum depth ($\mathrm{Max}_{\mathrm{Depth}}$), (2) when the number of segments in a leaf is smaller than a predefined threshold ($\mathrm{Min}_{\mathrm{Leaf}}$), or (3) when a node purely includes segments of one class. In recall (test) mode, the outputs of the RF are the seizure (p_s) and nonseizure (p_n) probabilities averaged from all outputs of decision trees.

In the proposed method, $\mathrm{Max}_{\mathrm{Depth}}$ and $\mathrm{Min}_{\mathrm{Leaf}}$ were respectively equal to $(Q-1)$ and 1, which means that if the third stopping condition is not reached, the tree can be as deep as possible, having one leaf for each bootstrapped training segment. As mentioned, 100 of these decision trees are trained over bootstrapped training segments with randomly selected features. For each test segment, the obtained seizure probability (p_s) is compared with a threshold to score the segment as seizure or nonseizure. This threshold is varied from 0 to 1 in order to construct the ROC curve.

3. Results

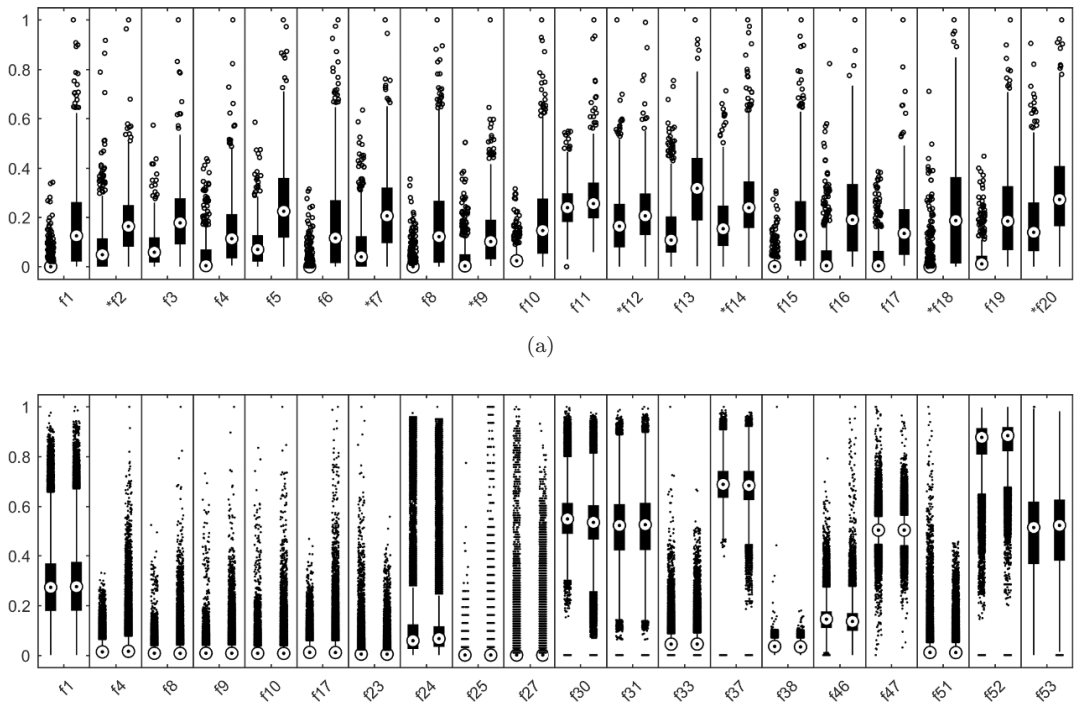
Figure 5 shows the box-plots of all extracted features by the (a) CNN and (b) Algorithm 3. In this figure, the features of Algorithm 3 are plotted because of its higher performance compared to Algorithms 1 and 2 (displayed in Fig. 7). For selecting the best 20 features of Algorithm 3, the LASSO method with 10-fold cross-validation was applied. The CNN features which were removed by the pruning process are marked with an asterisk ($*$). For each feature, the first and second boxes are corresponding to the nonseizure and seizure segments, respectively. In each plot, the filled black boxes show the first and third quartiles ($Q_{1,3}$) and the thin lines display the Whisker range from $Q_1 - 1.5 \times \text{IQR}$ to $Q_3 + 1.5 \times \text{IQR}$, where IQR is $Q_3 - Q_1$. The small circles in the plots show the outliers and big circles with a dot in the center show the median values. All features are plotted after being normalized between 0 and 1.

As mentioned, a bootstrap test was applied to compare different classifiers including LDA, SVM (linear and RBF kernels), and RF. Next, the AUC of each classifier was calculated when the features extracted by the CNN and Algorithm 3 are used



(a) (b)

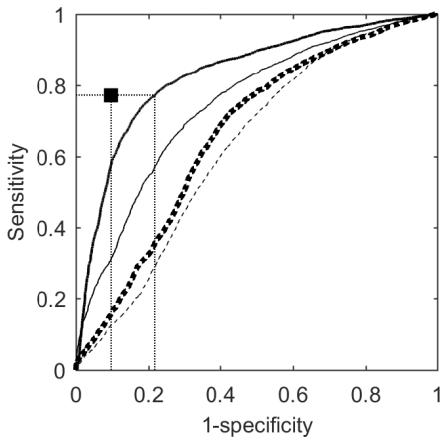
Fig. 6. The AUCs of different classifiers when the features are extracted by the (a) CNN and (b) Algorithm 3.



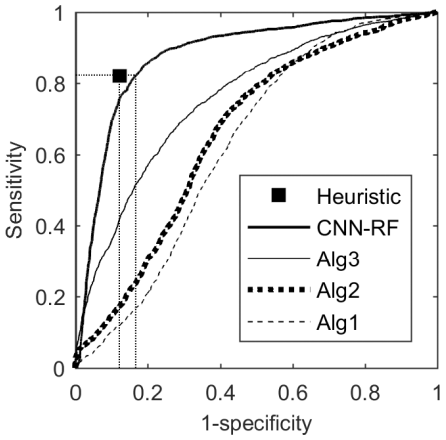
(a)

(b)

Fig. 5. Summary of the extracted features from the test dataset: (a) automatically extracted by the CNN. The features starting with star (*) are corresponding to the features removed after the pruning process. Panel (b) shows the selected features extracted in Algorithm 3 by a LASSO feature selection technique. All features are plotted after being normalized between 0 and 1.



(a)



(b)

Fig. 7. The ROC curves of the heuristic, feature-based, and proposed CNN–RF methods for the test data. (a) for all neonates in the test dataset and (b) after excluding seven neonates which did not have appropriate training patterns in the training dataset.

(Fig. 6). The Mann–Whitney–Wilcoxon statistical test applied on the results from RF versus SVM–RBF (* and ** in the figure) shows that the RF classifier has a significantly higher performance.

Figure 7 shows the ROC curves on the test dataset for the heuristic, feature-based algorithms (Algorithm 1–3), and the proposed CNN after pruning and connecting to the RF. The upper curve, Fig. 7(a), is the result of applying the proposed classifier to all test neonates (22 neonates), while the lower one, Fig. 7(b), shows the ROC when seven neonates, who expressed a completely different seizure pattern having no training patterns in the training dataset, were excluded from the training set. Results from the complete CNN, without pruning, are similar to the displayed CNN. The AUC of the CNN–RF method is also 8% higher than the pure CNN with the fully connected network (83% versus 75%).

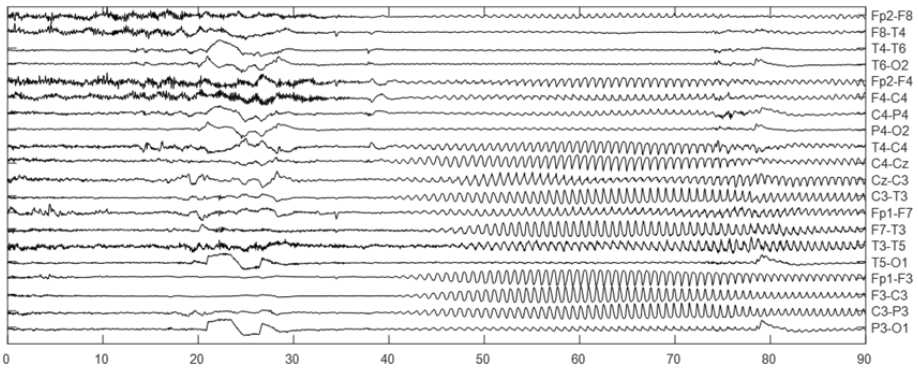
Table 3. Comparison of the performance metrics for the CNN and heuristic methods.

	Total database		After exclusion of seven neonates	
Metric	Heuristic	CNN	Heuristic	CNN
AUC (%)	88[a]	83	89[a]	88
Sensitivity (%)	77	77	82	82
Specificity (%)	90	78	88	84
GDR (%)	77	77	78	78
FAR (h^{-1})	0.63	0.90	0.77	0.73

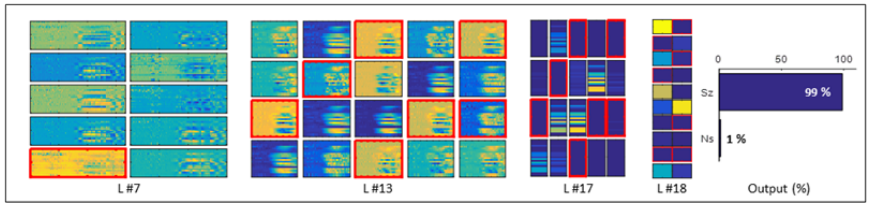
Note: [a]Using piecewise cubic Hermite interpolation.

Furthermore, in Table 3, the epoch-based AUC, sensitivity, and specificity, as well as event-based good detection rate (GDR) and false alarm rate per hour (FAR), are reported.[62,63] Since the output of the heuristic algorithm is not continuous, Hermite spline interpolation was used to calculate the AUC.[64] For other metrics, in order to make the comparison simpler, the threshold of the CNN–RF was chosen where the sensitivities of CNN–RF and heuristic methods are equal (the horizontal dashed lines in Fig. 7). As is clear from the table, after excluding the seven neonates, the specificity of CNN–RF is 5% less while the averaged false alarm rate per hour is 0.04 better than those of the heuristic methods. The results for individual neonates are displayed in Table B.1 (Appendix B) in detail.

Figures 8 and 9 show two qualitative examples of a seizure and a nonseizure segment, respectively. In these figures, the outputs of the seventh and eighth Conv layers, as well as the outputs of the 17th and 18th pooling layers, are shown. The red-highlighted images are corresponding to the filters that were removed by the pruning process. Each image of layers 7, 13, and 17 displays an output, which has 20 channels (y-axis), whereas the output of the layer 18, after pooling the channels, has only one value for all channels. Furthermore, as explained, the resolution of time (samples in x-axis in these figures) decreases after each pooling layer so that layers 17 and 18 have only one value in time. Therefore, the output of layer 18 includes 20 values (20 before pruning, and 13 after

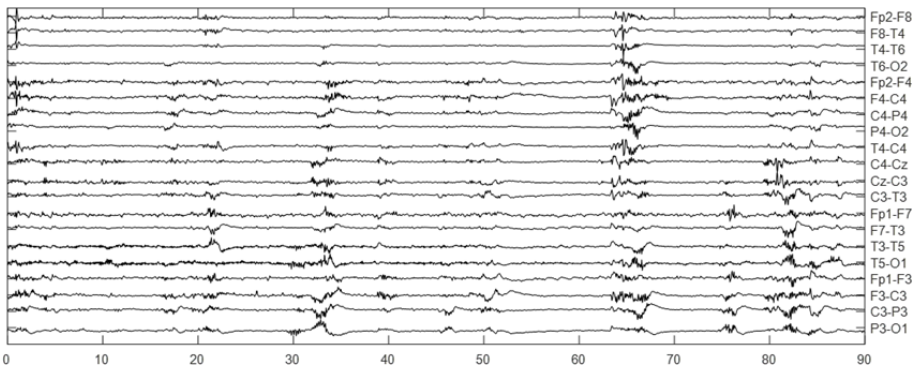


(a)

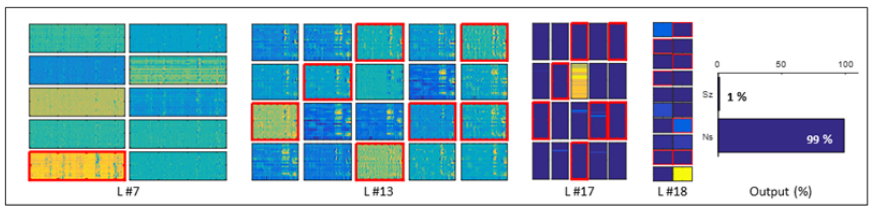


(b)

Fig. 8. (Color online) Qualitative example of a seizure segment and outputs for some layers. (a) A seizure segment with 20 bipolar channels. The x-axis is time in seconds. (b) The output of Conv layers 7 and 13, as well as pooling layers 17 and 18, and the final output of the CNN after the classification layers. The red-highlighted boxes correspond to the filters removed by the pruning process.



(a)



(b)

Fig. 9. (Color online) Qualitative example of a nonseizure segment and outputs for some layers. (a) A nonseizure segment with 20 bipolar channels. The x-axis is time in seconds. (b) The output of Conv layers 7 and 13, as well as pooling layers 17 and 18, and the final output of the CNN after the classification layers. The red-highlighted boxes correspond to the filters removed by the pruning process.

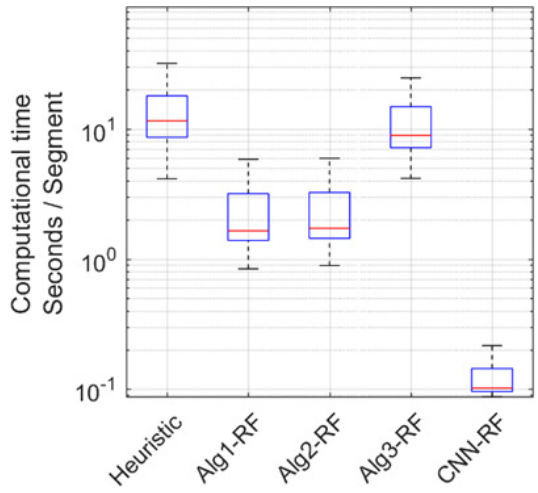


Fig. 10. The execution time is shorter for the CNN–RF when compared to the heuristic and feature-based methods. The time is measured in seconds for each segment (90 s, 20 channels).

pruning) each of which is a 1×1 (channel $\times$ time) number. These values are considered as the automatically extracted features. As is clear in the seizure example, Fig. 8, the seizure occurred in the right-hand (almost bottom) side of the segment. The outputs of the layer 7 show some variations in this area, which is different from the left-hand side of the segment. These differences are more pronounced in the outputs of the layer 13 where almost all filters were activated for the seizure area. For the layer 17, when the time information is completely compressed, the channels with maximum activation of some of the filters are clearly distinguishable. In contrast, for the nonseizure segment displayed in Fig. 9, there is no clear activation in the layers 7 and 13, and consequently no distinct channels in layer 17.

Figure 10 shows the recall (test phase) computational times calculated for the heuristic, feature-based (Algorithms 1–3), and proposed CNN–RF methods. The time is shown in seconds per each segment in a logarithmic scale. In order to have a correct comparison and overcome variable CPU loading, the methods ran chronologically for each segment and the elapsed times for each method and each segment were stored. Then, the median, Q_1 and Q_3 of the whole segments were calculated and shown in a boxplot. The time was measured from the moment of loading data to when the label was defined including preprocessing, feature extraction, and classification times. The algorithms ran in Matlab^{TM} platform, version 9.1.0 (2016b) (The MathWorks, Natick, MA, USA), and on a server computer, Intel(R) Xeon(R) CPU 2.20 GHz, with GNU/Linux operating system (Red Hat 4.8).

4. Discussion

In neonatal seizure detection using machine learning approaches, choosing a proper type of classifier plays an important role to have a good performance. Moreover, finding appropriate features that are discriminative and informative is another challenge and has a big influence on the performance. In neonatal seizure detection, like other classification problems, some researchers have discovered and proposed new features to enhance the performance, while others improved the classification strategies. By using a deep convolutional neural network, both the features as well as the classifier are optimized simultaneously. In this paper, a CNN with 18 layers was designed in order to automatically extract the required features from raw multi-channel EEG segment and a random forest was used to classify them. It is important to note that the final layers of the CNN method were also able to classify the segment based on the features extracted in the previous layers. However, they were replaced with a random forest, after training the network, in order to have a higher performance and a smoother ROC curve.

In the classic feature-based seizure detectors, like Algorithms 1–3, the features are hand-engineered and usually found by trial and error. It is possible that some information that is important for classification is fully or partly missed in the selected features. However, in the proposed method, since the feature extraction is optimized based on the training data, the maximum information, related to the classification, is potentially able to be extracted which improves the performance of the classification (compare CNN–RF and Algorithm 1–3 in Fig. 7). As shown in Fig. 5, the features extracted by the CNN are more discriminative than those extracted by Algorithm 3. This resulted from the optimization of feature learning process in the CNN, which is considered as the main advantage of the proposed method.

On average, the proposed CNN-based method performs better than the tested methods relying on features. However, as plotted in Fig. 7(a), the proposed method has lower performance than

the heuristic method due to the limited number of training neonatal EEG data sets. If a larger scored database was available for training, this could improve the performance of the method in a patient-independent framework. When lacking training data, it seems likely that heuristic methods, which are not directly trained based on the training data, will perform better. This is because of the fact that for developing the heuristic method mimicking an expert human EEG scorer, the knowledge of the neurophysiologist which has been collected over years from hundreds of neonates was used. It seems likely that to have a fair comparison between the heuristic method and data-driven techniques, adequate training data should be provided for the data-driven methods. If such a large training database is collected in future years, better methods could be developed. These methods have some important advantages, like retrainability, more accuracy, and more flexibility. In the test dataset, seven neonates with moderate to severe hypoxic brain injury were seen to have very unique seizure patterns. Similar patterns were not available in the training dataset. It is unlikely that a data-driven classifier will detect a pattern that was not presented in the training data. As it is shown in Table 3 and Fig. 7(b), if these seven neonates are excluded from the test dataset the performance of the proposed method is similar to the heuristic method and better than other data-driven methods. It is evident that by using a larger training dataset in the future, the performance of the trainable classifiers will be improved, while the heuristic "untrainable" methods will be unchanged. When the latent variable of the CNN method was fixed to provide the performance reported in Table 3, only 7% and 10% of seizure segments were detected by the CNN–RF method and by the heuristic method, respectively. This shows that even with the current lack of training data, the proposed CNN–RF method is a complementary method, which can detect 7% of seizures that were missed by the heuristic approach.

Since neonatal seizures are usually very focal or regional, which means that only one or a few channels display seizure patterns, retrospectively developed algorithms were designed as a single-channel detector so that they are applied on each EEG channel separately. If a seizure is detected in at least one channel, the segment is classified as seizure ("OR"). This idea is also applied in the proposed CNN method. As was explained, all filters in the Conv layers, as well as pooling layers, developed for extracting features, are $1 \times N$ operators. It means that they only affect the time mode and have no operation on the channel mode. However, it is likely that there is some spatial information, at least among adjacent electrodes, which show how seizures spread through channels in regional seizures, involving contiguous brain regions. Furthermore, each electrode is usually used more than once in different montage maps. For instance, in the full montage map, C4 is used four times in F4–C4, C4–P4, T4–C4, and C4–Cz. Hence, if a seizure occurs in a brain area close to the electrode C4, these four bipolar channels should display it. This spatial connection of channels can be very useful even for very focal seizures, and it can be a distinctive characteristic for distinguishing seizures from some artifacts. The CNN structure is easily able to extract this information by increasing the dimension of filters in the channel mode, as it is working in diverse image processing applications. However, since some electrodes were not recorded or available in our training database for some neonates and different bipolar montage maps were subsequently constructed, this information was not extractable. In case of datasets with homogeneous recordings, the filters can become two-dimensional and it is expected that the performance will be significantly improved. This is also true for Algorithm 1, which concatenates the features of different channels.

As previously explained, in Algorithm 2, the method is applied on each channel individually and an OR operator is then used to aggregate the channels. Since most neonatal seizures are regional/focal, the channel(s) representing seizure patterns should be predefined by the developer or an expert EEG reader for each training segment in order to train the method with correct training data. This is a very labor-intensive task especially in big datasets. This problem is exacerbated for Algorithm 3 where not only the true channels should be predefined, but also the proper epochs of those channels should be marked. Nevertheless, one of the main characteristics of the CNN is a shift-invariant property which means shifting the target pattern (like seizure pattern in this problem) through the first or second modes (time or channels here) has no effect on the output. This characteristic results from the parameter sharing of the neurons in the convolutional layers, as

well as from the maximum operation in the pooling layers. Therefore, for the proposed CNN method, it is not important in which channel or at what time the seizure activity is manifested. The CNN can automatically find the related region, so-called region-of-interest. To illustrate this, in the example of Fig. 8, where the seizure emerges in the almost bottom-right of the segment, the output of layer 13 shows that the CNN neurons were successfully activated in the seizure area, compared to Fig. 9 for the nonseizure segment.

Furthermore, in neonates, most seizures are rhythmic, with evolution of amplitude. For these seizures, the exact moment of onset or offset of seizures is sometimes not very clear due to the fact that the seizure patterns (oscillatory or spike-train) start with a very low amplitude and gradually increase over time (and vice versa for the offset). It is evident that even expert EEG readers may not agree with each other about the exact time and duration of seizures in this case.[55] Furthermore, in focal/regional seizures, the channels that are not close to the center of the seizure might show some low-amplitude seizure patterns, increasing the uncertainty for the classifier. Due to its shift-invariant property, the CNN method can overcome this inherent fuzzy onset and offset of seizures, as well as representing channels, so that it does not need to know when and where exactly the location of the seizure is within the segment.

In addition, retrospective studies of neonatal seizure detection have reported varying levels of agreement between expert EEG readers, with kappa coefficients ranging between 0.4 and 0.93.[13,55,65–68] It shows that different experts, and consequently different centers, use empirically different definitions of seizure and the gold standard is not yet clear-cut. One advantage of data-driven methods, like the proposed one, compared to fixed heuristic ones, which may work very well in one center, is that the network can be retrained or retuned in different centers in order to tailor to their needs. Moreover, an advantage of artificial neural networks, like the proposed CNN, is that different training segments can have different weights in the backpropagation training. Therefore, the segments upon which the experts have higher agreement can have more influence in the cost function during training than segments with a larger uncertainty. This weighted training technique can increase the overall satisfaction of the experts from the final outputs when the method is being used in different centers. Although the labels of only one expert reader were used in the present work, a multi-score analysis can be performed in the future.

Finally, one of the most important advantages of the proposed CNN–RF method is that it is made by simple FIR filters, maximum, and averaging operators. Consequently, the recall time is much faster than other tested heuristic and feature-based methods, see Fig. 10. This method is about 115, 89, and 17 times faster than the heuristic, Algorithm 3, and Algorithms 1 and 2, respectively, in recall. The computational time is very important for the real-time implementation. Although each of them can individually work in real-time and needs less than 90 s to process a 90 s segment, the real-time ratio (= average computational time needed for a segment/90 s) can be important when the final product acquires, filters, down-samples, stores, monitors the EEG, sends the data to a cloud system, and performs many other possible tasks and processes. Faster processing often results from less operational calculations, which means lower energy requirements, and it is very important in portable/wearable devices with limited source of energy.

However, the proposed approach has some disadvantages: first, this method, and in general all deep networks, is very time-consuming in the training process. Thus, optimizing the hyper-parameters or improving the design of the network is much harder than in a simple feature-based data-driven technique. Second, the suggested network, like other DNNs, needs a large amount of data to be trained. If sufficient training data are not available, it is very likely to over-fit due to the high number of layers and parameters. Third, compared to heuristic methods, the process is not transparent and it is not immediately evident why a certain segment is classified as seizure or nonseizure. Finally, compared to regular feature-based methods, the extracted features are just some numbers resulting from filtering, pooling, and rectifying of the EEG, which make it difficult to provide a tangible interpretation.

Several limitations of this study need to be acknowledged. First, the designed network and its structural parameters, including the number of layers, the length of filters, the strides, etc., have been chosen by trial and error and consequently they are not guaranteed to be optimal. Second, the scored

seizures used for training and testing were labeled by only one expert clinical neurophysiologist. Third, the data used in this paper were recorded in one center only. In order to have a more generalizable comparison, the methods should be tested on an extensive and multi-rated, multi-center database.

5. Conclusion

A novel neonatal seizure detector using convolutional neural networks and random forest was introduced in this paper. The main advantage of the proposed method is that it does not require hand-engineered feature extraction process, but it automatically extracts the required features and optimizes them based on the training data. We show that this proposed method outperforms the tested feature-based approaches. Compared to the previously developed heuristic detector, the proposed method is not yet superior because of the limited number of training neonates. However, it seems possible that by having more training data in future, it can reach the performance of the heuristic method as well. At last, it was also shown that the proposed method is remarkably faster than other tested algorithms, which is very important for real-time applications. However, further studies need to be carried out in order to validate this algorithm in a multi-center and multi-scored database. Furthermore, it seems that using sequence learners like hidden Markov models or LSTM instead of the used RF classifier can enhance the performance.

Acknowledgements

Amir H. Ansari and Sabine Van Huffel are supported by: Bijzonder Onderzoeksfonds KU Leuven (BOF): Center of Excellence (CoE) No. PFV/10/002 (OPTEC); SPARKLE: Sensor-based Platform for the Accurate and Remote monitoring of Kinematics Linked to E-health No. IDO-13-0358; "The effect of perinatal stress on the later outcome in preterm babies" No. C24/15/036; TARGID: Development of a novel diagnostic medical device to

Table A.1. Layers of the designed network after pruning.

		Layer Info	Output Size
Input:			(20, 2700, 1)
Feature Extraction	1	Conv$(1,5) \times \mathbf{3}$	(20, 2696, **3**)
	2	MPool$(1,3), s:2$	(20, 1347, **3**)
	3	ReLU	(20, 1347, **3**)
	4	Conv$(1,5) \times \mathbf{4}$	(20, 1343, **4**)
	5	MPool$(1,3), s:2$	(20, 671, **4**)
	6	ReLU	(20, 671, **4**)
	7	Conv$(1,5) \times 9$	(20, 667, **9**)
	8	MPool$(1,3), s:2$	(20, 333, **9**)
	9	ReLU	(20, 333, **9**)
	10	Conv$(1,5) \times 10$	(20, 329, **10**)
	11	MPool$(1,3), s:2$	(20, 164, **10**)
	12	ReLU	(20, 164, **10**)
	13	Conv$(1,20) \times 13$	(20, 145, **13**)
	14	MPool$(1,10), s:5$	(20, 28, **13**)
	15	ReLU	(20, 28, **13**)
	16	MPool$(1,5), s:3$	(20, 8, **13**)
	17	APool$(1,8), s:1$	(20, 1, **13**)
	18	MPool$(20,1), s:1$	(1, 1, **13**)
Classifier	19	Conv$(1,1) \times 5$	(1, 1, 5)
	20	Sigmoid	(1, 1, 5)
	21	Conv$(1,1) \times 2$	(1, 1, 2)
	22	Sigmoid	(1, 1, 2)
	23	Loss	(1, 1, 1)
		Total number of parameters: 3300	

assess gastric motility No. C32-16-00364; Fonds voor Wetenschappelijk Onderzoek Vlaanderen (FWO): Project No. G.0A5513N (Deep brain stimulation); Agentschap Innoveren & Ondernemen (VLAIO): Project STW 150466-OSA+ and O & O HBC 2016 0184 eWatch; IMEC: Strategic Funding 2017, No. ICON-HBC.2016.0167 SeizeIT; Belgian Federal Science Policy Office: IUAP No. 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, No. 316679; The HIP Trial: No. 260777; ERASMUS +: NGDIVS 2016-1-SE01-KA203-022114; European Research Council (ERC) Advanced Grant, No. 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. Alexander Caicedo and Amir H. Ansari are supported by IWT PHD Grant: TBM 110697-NeoGuard. Alexander Caicedo is a Postdoctoral Fellow from the Fonds voor Wetenschappelijk Onderzoek Vlaanderen (FWO).

Table B.1. The results of the CNN for each neonate.

	AUC (%)	True detection/ Total seizure	False alarm	Rec. len. (h)
1	93.8	14/14	7	4.0
2	92.6	8/9	2	4.0
3	88.3	5/6	0	4.0
4	88.1	2/2	5	2.0
5	84.8	2/2	0	2.0
6	83.6	8/9	8	4.0
7	83.4	39/43	4	4.0
8	83.0	16/18	6	4.0
9	82.8	30/30	6	4.1
10	82.2	4/6	0	4.0
11	82.0	4/5	12	4.0
12	77.7	5/22	2	4.0
13	77.6	12/14	0	2.0
14	76.6	27/29	7	4.0
15	76.4	8/18	0	2.0
16	74.0	13/15	0	2.1
17	72.5	2/14	0	4.0
18	72.0	3/13	0	4.0
19	71.4	11/23	0	2.1
20	65.9	28/28	3	4.0
21	62.9	2/9	1	2.0
22	60.9	43/44	4	4.0

Note: Rec. len. is the recording length in hours.

Appendix A. Pruned Network

Table A.1 lists the layers and parameters of the pruned CNN. The bold-faced values, in this table, show the different number of filters and the output size compared to Table 2.

Appendix B. Performance in Detail

Table B.1 shows the performance of the proposed CNN-based method for the tested neonates individually. In this table the AUC, the number of truly detected seizures, the total number of seizures, the number of false alarms, and the length of recordings are listed for each neonate.

References

1. J. M. Rennie *et al.*, Non-expert use of the cerebral function monitor for neonatal seizure detection, *Arch. Dis. Child. Fetal Neonatal Ed.* **89** (2004) F37–F40.
2. J. J. Volpe, *Volpe's Neurology of the Newborn* (Elsevier Health Sciences, 2017).
3. P. J. Cherian, *Improvements in Neonatal Brain Monitoring after Perinatal Asphyxia* (Erasmus University Rotterdam, 2010).
4. R. Pressler, Neonatal seizures, in *Introduction to Epilepsy* (Cambridge University press, 2012), pp. 142–149.
5. J. M. Rennie, Neonatal seizures, *Eur. J. Pediatr.* **156** (1997) 83–87.
6. D. M. Murray *et al.*, Defining the gap between electrographic seizure burden, clinical expression and staff recognition of neonatal seizures, *Arch. Dis. Child., Fetal Neonatal Ed.* **93** (2008) F187–F191.
7. A. Liu, J. S. Hahn, G. P. Heldt and R. W. Coen, Detection of neonatal seizures through computerized EEG analysis, *Electroencephalogr. Clin. Neurophysiol.* **82** (1992) 30–37.
8. J. Gotman, D. Flanagan, J. Zhang and B. Rosenblatt, Automatic seizure detection in the newborn: methods and initial evaluation, *Electroencephalogr. Clin. Neurophysiol.* **103** (1997) 356–362.
9. P. Celka and P. Colditz, A computer-aided detection of EEG seizures in infants: A singular-spectrum approach and performance comparison, *IEEE Trans. Biomed. Eng.* **49** (2002) 455–462.
10. M. A. Navakatikyan *et al.*, Seizure detection algorithm for neonates based on wave-sequence analysis, *Clin. Neurophysiol.* **117** (2006) 1190–1203.
11. W. Deburchgraeve *et al.*, Automated neonatal seizure detection mimicking a human observer reading EEG, *Clin. Neurophysiol.* **119** (2008) 2447–2454.
12. M. De Vos *et al.*, Automated artifact removal as preprocessing refines neonatal seizure detection, *Clin. Neurophysiol.* **122** (2011) 2345–2354.

13. P. J. Cherian *et al.*, Validation of a new automated neonatal seizure detection system: A clinician's perspective, *Clin. Neurophysiol.* **122** (2011) 1490–1499.
14. H. Hassanpour, M. Mesbah and B. Boashash, Time–frequency based newborn EEG seizure detection using low and high frequency signatures, *Physiol. Meas.* **25** 935 (2004).
15. B. R. Greene *et al.*, A comparison of quantitative EEG features for neonatal seizure detection, *Clin. Neurophysiol.* **119** (2008) 1248–1261.
16. E. M. Thomas, A. Temko, G. Lightbody, W. P. Marnane and G. B. Boylan, Gaussian mixture models for classification of neonatal seizures using EEG, *Physiol. Meas.* **31** (2010) 1047–1064.
17. R. Ahmed, A. Temko, W. P. Marnane, G. Boylan and G. Lightbody, Exploring temporal information in neonatal seizures using a dynamic time warping based SVM kernel, *Comput. Biol. Med.* **82** (2017) 100–110.
18. A. Temko, E. Thomas, W. Marnane, G. Lightbody and G. Boylan, EEG-based neonatal seizure detection with support vector machines, *Clin. Neurophysiol.* **122** (2011) 464–473.
19. S. B. Nagaraj, N. J. Stevenson, W. P. Marnane, G. B. Boylan and G. Lightbody, Neonatal seizure detection using atomic decomposition with a novel dictionary, *IEEE Trans. Biomed. Eng.* **61** (2014) 2724–2732.
20. A. Zwanenburg *et al.*, Using trend templates in a neonatal seizure algorithm improves detection of short seizures in a foetal ovine model, *Physiol. Meas.* **36** (2015) 369.
21. A. Aarabi, F. Wallois and R. Grebe, Automated neonatal seizure detection: A multistage classification system through feature selection based on relevance and redundancy analysis, *Clin. Neurophysiol.* **117** (2006) 328–340.
22. J. Mitra *et al.*, A multi-stage system for the automated detection of epileptic seizures in neonatal EEG, *J. Clin. Neurophysiol.* **26** (2009) 218–226.
23. A. H. Ansari *et al.*, Improved multi-stage neonatal seizure detection using a heuristic classifier and a data-driven post-processor, *Clin. Neurophysiol.* **127** (2016) 3014–3024.
24. Y. Bengio, Learning deep architectures for AI, *Found. Trends Mach. Learn.* **2** (2009) 1–127.
25. S. Ghosh-Dastidar and H. Adeli, A new supervised learning algorithm for multiple spiking neural networks with application in epilepsy and seizure detection, *Neural Netw.* **22** (2009) 1419–1431.
26. S. Ghosh-Dastidar and H. Adeli, Improved spiking neural networks for EEG classification and epilepsy and seizure detection, *Integr. Comput.-Aided Eng.* **14** (2007) 187–212.
27. S. Ghosh-Dastidar, H. Adeli and N. Dadmehr, Principal component analysis-enhanced cosine radial basis function neural network for robust epilepsy and seizure detection, *IEEE Trans. Biomed. Eng.* **55** (2008) 512–518.
28. W. R. S. Webber, R. P. Lesser, R. T. Richardson and K. Wilson, An approach to seizure detection using an artificial neural network (ANN), *Electroencephalogr. Clin. Neurophysiol.* **98** (1996) 250–272.
29. A. J. Gabor, R. R. Leach and F. U. Dowla, Automated seizure detection using a self-organizing neural network, *Electroencephalogr. Clin. Neurophysiol.* **99** (1996) 257–266.
30. Y. LeCun and Y. Bengio, Convolutional networks for images, speech, and time series, in *The Handbook of Brain Theory and Neural Network* (MIT Press, 1998), pp. 255–258.
31. G. E. Hinton, S. Osindero and Y.-W. Teh, A fast learning algorithm for deep belief nets, *Neural Comput.* **18** (2006) 1527–1554.
32. P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio and P.-A. Manzagol, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, *J. Mach. Learn. Res.* **11** (2010) 3371–3408.
33. S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Comput.* **9** (1997) 1735–1780.
34. B. Hutchinson, L. Deng and D. Yu, Tensor deep stacking networks, *IEEE Trans. Pattern Anal. Mach. Intell.* **35** (2013) 1944–1957.
35. O. Abdel-Hamid, A. Mohamed, H. Jiang and G. Penn, Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition, in *Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2012), pp. 4277–4280.
36. D. C. Ciresan, U. Meier, J. Masci, L. Maria Gambardella and J. Schmidhuber, Flexible, high performance convolutional neural networks for image classification, in *Proc. 22nd Int. Joint Conf. Artificial Intelligence*, Vol. 2 (AAAI press, 2011), pp. 1237–1242.
37. S. Lawrence, C. L. Giles, A. C. Tsoi and A. D. Back, Face recognition: A convolutional neural-network approach, *IEEE Trans. Neural Netw.* **8** (1997) 98–113.
38. K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556 [cs.CV].
39. A. Krizhevsky, I. Sutskever and G. E. Hinton, Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems* (NIPS Foundation, 2012), pp. 1097–1105.
40. H. Cecotti and A. Graeser, Convolutional neural network with embedded fourier transform for EEG classification, in *Proc. 19th Int. Conf. Pattern Recognition* (IEEE, 2008), pp. 1–4.
41. H. Cecotti and A. Graeser, Convolutional neural networks for P300 detection with application

to brain-computer interfaces, *IEEE Trans. Pattern Anal. Mach. Intell.* **33** (2011) 433–445.
42. P. W. Mirowski, Y. LeCun, D. Madhavan and R. Kuzniecky, Comparing SVM and convolutional networks for epileptic seizure prediction from intracranial EEG, in *Proc. IEEE Workshop Machine Learning for Signal Processing* (IEEE, 2008), pp. 244–249.
43. U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan and H. Adeli, Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals, *Comput. Biol. Med.* (2017), doi:10.1016/j.compbiomed.2017.09.017.
44. A. O'Shea, G. Lightbody, G. Boylan and A. Temko, Neonatal seizure detection using convolutional neural networks, arXiv:1709.05849 [Stat.ML].
45. T. N. Sainath, O. Vinyals, A. Senior and H. Sak, Convolutional, long short-term memory, fully connected deep neural networks, in *Proc. 2015 IEEE Int. Conf. Acoustics, Speech and Signal Processing* (*ICASSP*) (IEEE, 2015), pp. 4580–4584.
46. M. Tan, C. dos Santos, B. Xiang and B. Zhou, LSTM-based deep learning models for non-factoid answer selection, arXiv:1511.04108 [Cs.CL].
47. S. Bai, Growing random forest on deep convolutional neural networks for scene categorization, *Expert Syst. Appl.* **71** (2017) 279–287.
48. X.-X. Niu and C. Y. Suen, A novel hybrid CNN–SVM classifier for recognizing handwritten digits, *Pattern Recognit.* **45** (2012) 1318–1325.
49. A. Sharif Razavian, H. Azizpour, J. Sullivan and S. Carlsson, CNN features off-the-shelf: An astounding baseline for recognition, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops* (2014), pp. 806–813.
50. A. Temko, G. Boylan, W. Marnane and G. Lightbody, Robust neonatal EEG seizure detection through adaptive background modeling, *Int. J. Neural Syst.* **23** (2013) 1350018.
51. A. H. Ansari *et al.*, Improved neonatal seizure detection using adaptive learning, in *Proc. 2017 39th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society* (*EMBC*) (2017), pp. 2810–2813.
52. V. Apgar, A proposal for a new method of evaluation of the newborn, *Curr. Res. Anesth. Analg.* **32**(4) (1953) 260–267.
53. V. Apgar, D. A. Holaday, L. S. James, I. M. Weisbrot and C. Berrien, Evaluation of the newborn infant-second report, *J. Am. Med. Assoc.* **168** (1958) 1985–1988.
54. H. B. Sarnat and M. S. Sarnat, Neonatal encephalopathy following fetal distress: A clinical and electroencephalographic study, *Arch. Neurol.* **33** (1976) 696–705.
55. A. Dereymaeker *et al.*, Interrater agreement in visual scoring of neonatal seizures based on majority voting on a web-based system: The Neoguard EEG database, *Clin. Neurophysiol.* **128** (2017) 1737–1745.
56. P. J. Cherian, R. M. Swarte and G. H. Visser, Technical standards for recording and interpretation of neonatal electroencephalogram in clinical practice, *Ann. Indian Acad. Neurol.* **12** (2009) 58–70.
57. S. R. Mathieson *et al.*, Validation of an automated seizure detection algorithm for term neonates, *Clin. Neurophysiol.* **127** (2016) 156–168.
58. A. H. Ansari *et al.*, Improvement of an automated neonatal seizure detector using a post-processing technique, in *Proc. 2015 37th Ann. Int. Conf. IEEE Engineering in Medicine and Biology Society* (*EMBC*) (IEEE, 2015), pp. 5859–5862.
59. S. Faul, G. Boylan, S. Connolly, W. Marnane and G. Lightbody, Chaos theory analysis of the newborn EEG: Is it worth the wait? in *Proc. IEEE Int. Workshop Intelligent Signal Processing* (2005), pp. 381–386.
60. A. Vedaldi and K. Lenc, MatConvNet: Convolutional neural networks for MATLAB, in *Proc. ACM Int. Conf. Multimedia* (2015).
61. L. Breiman, J. Friedman, C. J. Stone and R. A. Olshen, *Classification and Regression Trees* (CRC Press, 1984).
62. A. Temko, E. Thomas, W. Marnane, G. Lightbody and G. B. Boylan, Performance assessment for EEG-based neonatal seizure detectors, *Clin. Neurophysiol.* **122** (2011) 474–482.
63. A. H. Ansari *et al.*, Weighted performance metrics for automatic neonatal seizure detection using multi-scored EEG data, *IEEE J. Biomed. Health Inform.* (2017), doi:10.1109/JBHI.2017.2750769.
64. F. Fritsch and R. Carlson, Monotone piecewise cubic interpolation, *SIAM J. Numer. Anal.* **17** (1980) 238–246.
65. D. K. Shah *et al.*, Accuracy of bedside electroencephalographic monitoring in comparison with simultaneous continuous conventional electroencephalography for seizure detection in term infants, *Pediatrics* **121** (2008) 1146–1154.
66. N. S. Abend *et al.*, Interobserver reproducibility of electroencephalogram interpretation in critically ill children, *J. Clin. Neurophysiol.* **28** (2011) 15–19.
67. C. J. Wusthoff *et al.*, Interrater agreement in the interpretation of neonatal electroencephalography in hypoxic-ischemic encephalopathy, *Epilepsia* **58** (2017) 429–435.
68. N. J. Stevenson *et al.*, Interobserver agreement for neonatal seizure detection using multichannel EEG, *Ann. Clin. Transl. Neurol.* **2** (2015) 1002–1011.