KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Ansari A. H., De Wel O., Lavanga M., Caicedo A., Dereymaeker A., Jansen K., Vervisch J., De Vos M., Naulaers G., Van Huffel S. (2018), **Quiet Sleep Detection in Preterm Infants using Deep Convolutional Neural Networks** Journal of Neural Engineering, vol. 15, nr 6, 066006, 2018, doi: 10.1088/1741-2552/aadc1f
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	insert link to the published version of your paper https://iopscience.iop.org/article/10.1088/1741-2552/aadc1f/meta
Journal homepage	insert link to the journal homepage of your paper https://iopscience.iop.org/journal/1741-2552
Author contact	your email ofelie.dewel@kuleuven.be your phone number *+ 32 (0)16 326931*
Abstract	Neonates spend most of their time asleep. Sleep of preterm infants evolves rapidly throughout maturation and plays an important role in brain development. Since visual labelling of the sleep stages is a time consuming task, automated analysis of the electroencephalography (EEG) to identify sleep stages is of great interest to clinicians. This automated sleep scoring can aid in optimizing neonatal care and assessing brain maturation. In this study, we designed and implemented an 18-layer convolutional neural network to discriminate quiet sleep from non-quiet sleep in preterm infants. The network is trained on 54 recordings from 13 preterm neonates and the performance is assessed on 43 recordings from 13 independent

	patients. All neonates had a normal neurodevelopmental outcome and the EEGs were recorded between 27 and 42 weeks postmenstrual age. The proposed network achieved an area under the mean and median ROC curve equal to 92% and 98% respectively.
IR	url in Lirias

Quiet Sleep Detection in Preterm Infants using Deep Convolutional Neural Networks

Amir Hossein Ansari[†,1,2], Ofelie De Wel[†,1,2], Mario Lavanga[1,2], Alexander Caicedo[1,2], Anneleen Dereymaeker[3], Katrien Jansen[3,4], Jan Vervisch[3], Maarten De Vos[5], Gunnar Naulaers[3] and Sabine Van Huffel[1,2]

[1] Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium
[2] imec, Leuven, Belgium
[3] Department of Development and Regeneration, University Hospitals Leuven, Neonatal Intensive Care Unit, KU Leuven, Leuven , Belgium
[4] Department of Development and Regeneration, University Hospitals Leuven, Neonatal Intensive Care Unit & Child Neurology, KU Leuven, Leuven , Belgium
[5] Institute of Biomedical Engineering (IBME), Department of Engineering Science, University of Oxford, United Kingdom

E-mail: ofelie.dewel@kuleuven.be

February 2018
[†] *These authors are joint first authors*

Abstract. Neonates spend most of their time asleep. Sleep of preterm infants evolves rapidly throughout maturation and plays an important role in brain development. Since visual labelling of the sleep stages is a time consuming task, automated analysis of the electroencephalography (EEG) to identify sleep stages is of great interest to clinicians. This automated sleep scoring can aid in optimizing neonatal care and assessing brain maturation. In this study, we designed and implemented an 18-layer convolutional neural network to discriminate quiet sleep from non-quiet sleep in preterm infants. The network is trained on 54 recordings from 13 preterm neonates and the performance is assessed on 43 recordings from 13 independent patients. All neonates had a normal neurodevelopmental outcome and the EEGs were recorded between 27 and 42 weeks postmenstrual age. The proposed network achieved an area under the mean and median ROC curve equal to 92% and 98% respectively.

Keywords: Convolutional neural network; EEG; Sleep stage classification; Preterm neonate. Submitted to: *J. Neural Eng.*

1. Introduction

Preterm infants are born during a critical period of rapid growth and development of the brain and central nervous system. In these newborns, the fast neural growth that normally occurs in the womb during the third trimester of gestation is interrupted and has to take place in the neonatal intensive care unit (NICU). Therefore, a primary concern of NICUs is to support optimal brain development of these vulnerable infants, who are at increased risk of neurodevelopmental disorders [1, 2]. Continuous electroencephalography (EEG) monitoring is considered a valuable noninvasive tool to assess and track brain maturation of newborns in the NICU. Sleep wake cycling (SWC) undergoes fast changes and is one of the main hallmarks of neurodevelopment in preterm infants. Since sleep state organization reflects the level of functional brain development, a first step towards automated analysis of cerebral function is to identify the different sleep stages [3, 4].
Moreover, evidence suggests that sleep is of utmost importance in the development and maturation of neural pathways. As a consequence, the neonate's sleep should be protected and promoted during their stay in the NICU [5, 6]. Automated algorithms for real-time neonatal sleep state identification could optimise the scheduling of neonatal interventions and care procedures in the NICU. Up to now, the gold standard for the assessment of sleep EEG is visual analysis by an expert clinician. However, visual interpretation of neonatal EEG is a tedious and time consuming task for the clinician and requires expertise that might not always be available in the NICU [4]. This can be overcome by developing automated algorithms to discriminate neonatal vigilance states.

A growing body of literature has investigated algorithms for neonatal EEG sleep scoring [7–9], which typically rely on the difference in continuity [4, 7] and frequency content [10] between neonatal sleep states. Palmu et al. [11] compared the proportional duration of bursts, called Spontaneous Activity Transients (SAT%), in the EEG between neonatal sleep states. They confirmed the existence of genuine vigilance states in early preterm infants and found that the fluctuation of the SAT% signal corresponds to sleep stage cycling. Stevenson et al. [12] continued this work and derived quantitative measures of brain activity cycling from the frequency domain representation of the SAT% time series. Besides analysis of the SAT% signal, EEG time profiles have been used for automated sleep stage detection as well [13–15]. The EEG time profiles are constructed by adaptive segmentation of the EEG and subsequent clustering of features extracted from each segment. Thresholding of the processed time profile, which shows the class membership of the EEG segments over time, will then lead to labels of the sleep states. This concept of EEG time profiles for sleep stage modelling was explored by Barlow et al. [16], extended by Krajca et al. [13, 14] and further improved by Dereymaeker et al. [15]. In addition to the approaches described above, a number of studies have developed a neonatal sleep stage classifier based on a set of discriminative EEG features in combination with a machine learning classifier [7, 17–19].

While a wide range of features have been used to develop neonatal sleep stage

classifiers, there is still no consensus on the optimal combination of features. Moreover, those hand-crafted features are based on prior human knowledge about EEG sleep patterns, hence only a limited set of features have been examined. In general, the process of selecting the optimal feature set and classifier is a challenging task.

This study will address these problems by adopting a deep convolutional neural network (CNN) to detect quiet sleep in preterm infants. Deep neural networks learn the features directly from the data, which eliminates the need for manual feature extraction. CNNs have been extensively used in the field of image and speech processing [20–25]. Only recently, deep neural networks have been attracting a lot of interest for biomedical applications, such as medical image analysis [26], detection of myocardial infarction [27] and EEG seizure detection [28]. Furthermore, we have proposed a CNN for neonatal seizure detection in a previous study [29]. CNNs have been applied for sleep stage scoring in adults [30–35], however to the best of our knowledge this is the first paper adopting CNNs for neonatal sleep stage classification.

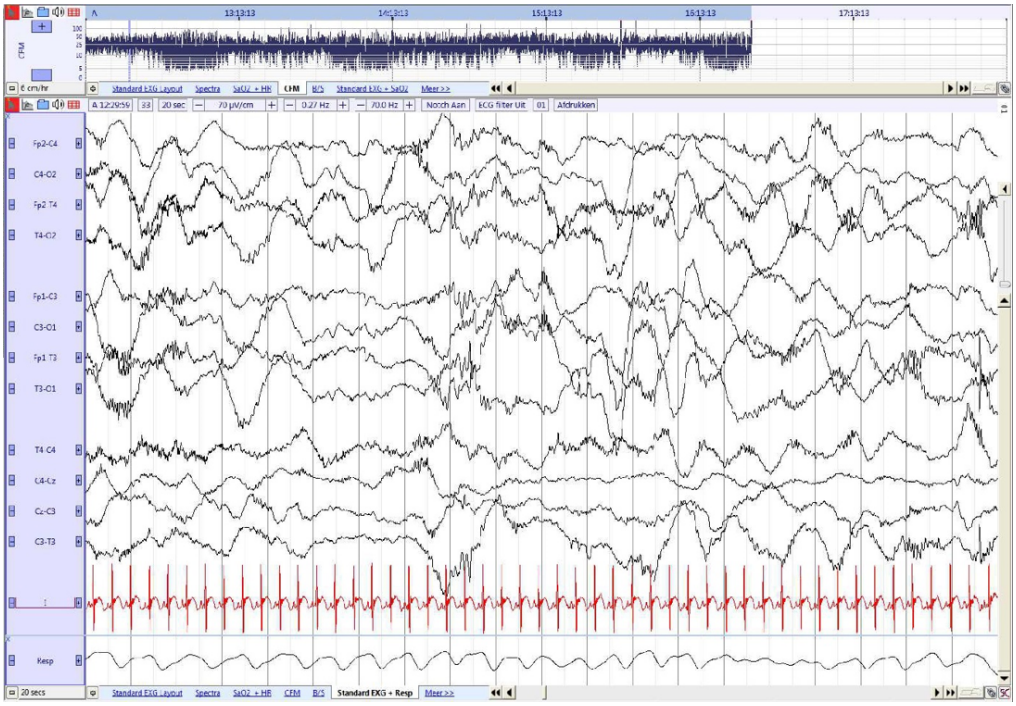
The aim of this paper is to design and train a deep convolutional neural network to identify sleep stages in preterm infants. The performance of this novel classifier will be compared to two reference algorithms described in the literature: (1) a support vector machine (SVM) classifier using a set of spectral features [10] and (2) CLuster-based Adaptive Sleep Staging (CLASS) developed by Dereymaeker et al. [15].

The remaining part of the paper proceeds as follows. We start by describing the database that has been used to develop, train and test the classifier. We will then go on to a short introduction to convolutional neural networks and a thorough description of the architecture of the network developed for sleep scoring. The pipeline of the two reference algorithms described in the literature will be briefly explained as well. Next, the performance of the proposed CNN and two reference algorithms will be reported and compared. At last, the advantages and disadvantages of the proposed method and future directions will be discussed.

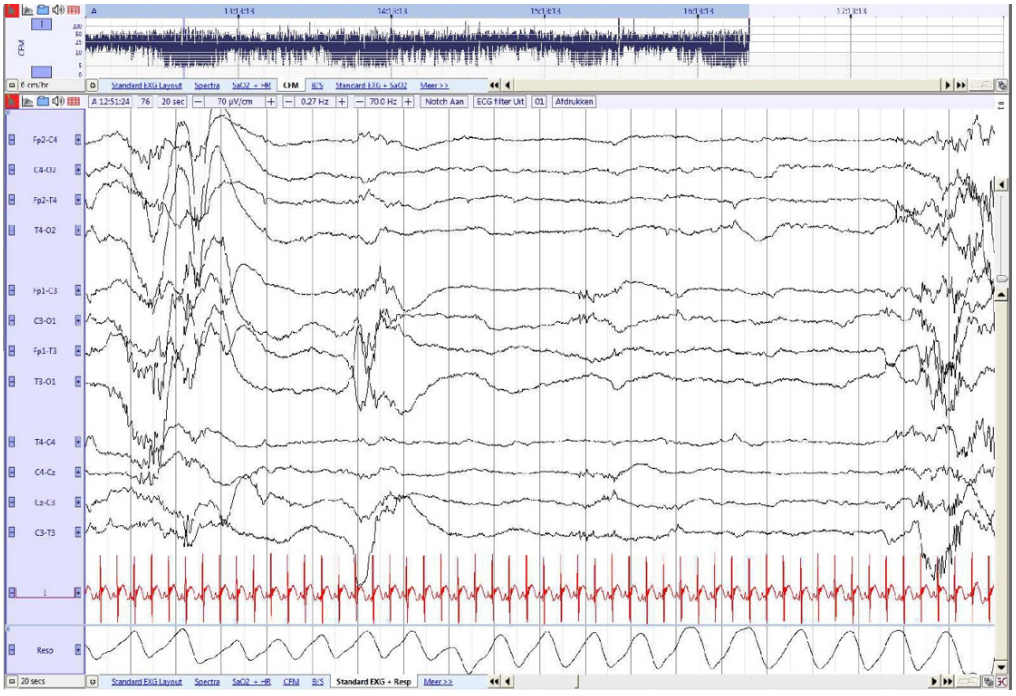
2. Materials and Methods

2.1. Database

The data used in this study were collected between 2012 and 2014 at the neonatal intensive care unit (NICU) of the University Hospitals Leuven, Belgium. The EEG signals were recorded after approval by the Ethics Committee of University Hospitals Leuven and informed parental consent were obtained. Serial EEG recordings collected from 26 preterm infants born before 32 weeks of gestation were used in this study. The EEG of each baby is recorded at least twice during their stay in the NICU, resulting in 97 multichannel EEG recordings between 27 and 42 weeks postmenstrual age (PMA). All neonates enrolled in the study had a normal neurodevelopmental outcome score at 9 and 24 months corrected age (Bayley Scales of Infant Development-II, mental and motor score > 85). Moreover, none of the subjects were under sedative or anti-



(a)



(b)

Figure 1. Example of a non-quiet sleep and quiet sleep EEG segment at 32 weeks and 2 days PMA. (a) Continuous tracing during active sleep. Delta brushes in temporal and occipital regions, irregular breathing pattern. (b) Discontinuous tracing during quiet sleep. IBI shorter than 15 s, temporal theta activity and occipital delta brushes, more regular breathing pattern.

epileptic medication during the EEG registration or had severe cerebral lesions (normal cerebral ultrasonography or intraventricular haemorrhage grade $\leqslant$ II, no periventricular leukomalacia or ventricular dilatation $>$ p97). The EEG was recorded using 9 electrodes: F1, F2, C3, C4, T3, T4, O1, O2 and reference electrode Cz, according to the modified international 10-20 system [36]. The monopolar EEG set up was used and the reference electrode Cz was discarded during the analysis phase. BrainRT equipment (OSG bvba, Rumst, Belgium) was employed to acquire the data. The data were initially filtered between 0.3 and 70 Hz and sampled at 250 Hz. The quiet sleep segments were annotated by two independent expert clinicians upon consensus. Wakefulness and other sleep stages (active sleep and indeterminate sleep) are merged and labelled as non-quiet sleep. An example of a quiet sleep and non-quiet sleep EEG segment is shown in Figure 1.
The data from 26 preterm infants were split into a training set and a test set. The training set containing 54 recordings from 13 patients was used to develop and train the convolutional neural network. The total duration of training data is equal to 269 h, with 69 h of quiet sleep, and 200 h of non-quiet sleep. The test set consisting of 43 recordings from 13 independent patients was used to assess the classification performance. The test set consists of 223 h of EEG with 53 h during quiet sleep and 170 h during non-quiet sleep. This interpatient data splitting avoids that the model is tested on data from a patient that is also used to train the model. In this way, we expect that patient specific characteristics can not bias the classification performance.

2.2. Convolutional neural networks

A convolutional Neural network (CNN), also known as ConvNet, is a special type of artificial feedforward neural network (ANN). These networks are biologically inspired by the processing of the mammalian visual cortex in which each neuron is stimulated by only a restricted region of the visual field, known as the receptive field [37–39]. CNNs usually have a deeper structure compared to ANNs (e.g. multilayer perceptrons, radial basis function networks), because they need fewer parameters for the same number of hidden layers. This parameter reduction, which is achieved by 3 key features of CNN: local receptive fields, parameter sharing and pooling, makes the computation more efficient and helps preventing overfitting. Using local receptive fields means that each neuron is only connected to a limited region of the input volume, in contrast with fully connected neurons. As a result, the network is able to learn local spatial/temporal correlations from the input. Parameter sharing means that all neurons located in one layer have the same parameters. This allows the network to detect a shifted target pattern, which is known as the shift invariant property. The last characteristic that makes the CNN different from other ANNs is pooling, which is a nonlinear down-sampler. Pooling progressively reduces the size of the data representation and improves the translational invariance.

Convolutional neural networks are comprised of 3 important layers: a convolutional layer (Conv), a rectified linear unit (ReLU) layer and a pooling layer. The input of the CNN as well as the output of each layer are usually 2 or 3-dimensional tensors dependent

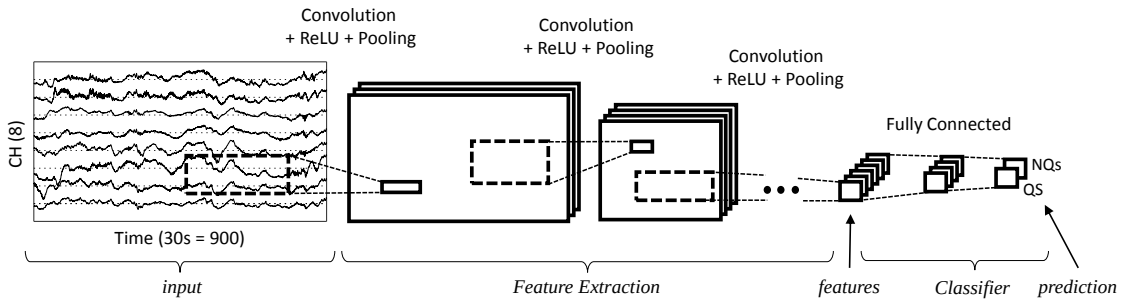


Figure 2. Architecture of the convolutional neural network.

on the input size and the architecture of the network. Each Conv layer typically consists of different feature maps and the output of its neurons is computed as

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

where k is the index of the feature map, I(i,j,p) is the input of the Conv layer, b_k is the bias of the k^{th} feature map, f_k (m,n,p) is the filter that should be convolved with the input in the first and second mode. So, a 2D convolution is performed. This filter consists of M $\times$ N $\times$ P coefficients, where M and N represent the size of the filter, while P represents the number of feature maps in the previous layer. O(i,j,k) is the output of convolutional layer and subsequently the input of the next layer. The filter coefficients, f_k, are unknown parameters that should be optimized in the backpropagation method. Besides, filter size (M,N), known as the receptive field, as well as the number of feature maps in each Conv layer are the hyper-parameters of the network and should be selected during the design process. For example, in the 6^{th} layer of the proposed method (Table 1), the convolutional layer is composed of 5 filters, each with size 1 $\times$ 5 (M=1, N=5). The ReLU activation function is a simple half-wave rectifier and operates as

$$y = max(0, x), \tag{2}$$

where x and y are the input and output of the unit respectively.
The pooling layer is a down-sampler, which uses a maximum or mean operator and has two parameters: size and stride. The former defines the spatial extent of the data that will be aggregated and the latter determines the shift which corresponds to the reduction rate. For instance, in the pooling layer in the 5^{th} layer of the proposed method (see Table 1), a window with size (2×3) is replaced by its maximum value and this window is shifted 2 samples in each mode (stride $= 2 \times 2$).
A CNN is composed of a sequence of the aforementioned layers, which extract the abstractions, and is then usually followed by a fully connected network with 1 or 2 hidden layers to perform classification.

Table 1. Layers of the designed network.

	Layer	Type	Size	No. filters	No. parameters	Stride	Padding	Output dimension
Feature Extraction	0	Input						(8,900,1)
	1	Conv	(1,10)	3	33	(1,1)	(0,9)	(8,900,3)
	2	ReLU						(8,900,3)
	3	Conv	(3,1)	3	30	(1,1)		(6,900,3)
	4	ReLU						(6,900,3)
	5	Mpool	(2,3)			(2,2)	(0,2)	(3,450,3)
	6	Conv	(1,5)	5	80	(1,1)	(0,4)	(3,450,5)
	7	ReLU						(3,450,5)
	8	Conv	(3,1)	5	80	(1,1)		(1,450,5)
	9	ReLU						(1,450,5)
	10	Mpool	(1,5)			(1,3)	(0,2)	(1,150,5)
	11	Conv	(1,5)	7	182	(1,1)	(0,7)	(1,150,7)
	12	ReLU						(1,150,7)
	13	Mpool	(1,6)			(1,4)		(1,37,7)
	14	Conv	(1,37)	10	2600	(1,1)		(1,1,10)
FC	15	Sigmoid						(1,1,10)
	16	Conv	(1,1)	2	22	(1,1)		(1,1,2)
	17	Sigmoid						(1,1,2)
	18	Softmax						(1,1,1)
Total Number of Parameters:					3027			

Conv: Convolutional layer
ReLU: Rectified linear unit
MPool: Pooling by maximum operator
FC: Fully connected classifier

2.3. The proposed CNN for sleep stage classification

Prior to using the EEG data as input for the CNN, some pre-processing steps are performed. First of all, the EEG is bandpass filtered between 1 and 15 Hz. The filtered EEG is then downsampled to 30 Hz in order to reduce the complexity of the network. The training data is normalized so that the mean and standard deviation of each EEG channel across the whole data set is equal to zero and one respectively. The parameters from the training data are then used to normalize the test data. At last, the EEG is segmented into windows of length 30 s, resulting in EEG segments of size 8 (channels) $\times$ 900 (samples).

These pre-processed EEG segments are then fed into a 18-layer CNN, whose

architecture is shown in Figure 2 and Table 1. The architecture of the network was designed based on our previous experience for neonatal seizure detection [29], and trial and error using the training and validation data. The CNN is designed and implemented in Matlab using the MatConvNet toolbox [40]. The first 14 layers perform feature extraction and the last 4 layers including the fully connected layers perform classification. The mask (kernel) size, stride, padding, and the number of feature maps used in the conv and/or pooling layers are also mentioned in the table.
For training the CNN using the backpropagation algorithm, 20% of the training data was selected as validation set and the remaining 80% of the training data was used to train the network's parameters. During this splitting of training data into training and validation set, the ratio of quiet sleep to non-quiet sleep segments is retained. The learning rate changes from 10e-2 to 10e-4 dependent on the training epoch and the layer. The weight decay was equal to 10e-6. For initializing the network, all bias terms were initiated with zero and the filter coefficients were set with Gaussian random noise with mean and standard deviation equal to 0 and 0.4 respectively. Batch learning was adopted with the batch size equal to 20 EEG segments and the maximum number of training epochs equal to 1000. In order to avoid overfitting and ensure network generalization early stopping was used. So, the validation set error was monitored during the training phase and training was stopped at epoch 493 which has the lowest validation set error.

According to the literature a sleep stage should last at least three consecutive minutes or three out of four consecutive minutes [41]. Relying on this fact, a moving average filter with length 6 (6 segments of 30 s = 3 min) is used as a post-processing step.
The proposed method will be compared to two methods from the literature, which will be briefly explained in the next subsections.

2.4. Feature-based neonatal sleep stage classifier

This neonatal sleep stage classifier is based on a set of spectral features described by Piryatinskaa et al. [10] which are fed into a support vector machine (SVM) classifier.

After bandpass filtering between 1 and 20 Hz and segmenting the multichannel EEG in epochs of 30 s, 9 spectral features are extracted from each EEG segment: (1-4) relative power in 4 EEG frequency bands (δ: 0.5 - 4 Hz, θ: 4 - 8 Hz, α: 8 - 12 Hz, β: 12 - 15 Hz), (5-6) spectral edge frequency (75% and 90%), (7) spectral moment, (8) spectral entropy and (9) amplitude entropy. Each multichannel EEG segment is then characterized by a total of 72 (9×8 channels) features and is marked as quiet sleep or non-quiet sleep according to the clinical label. The hyperparameters of the SVM with radial basis function (RBF) kernel are set using 5-fold cross validation on the training data. This SVM classifier is then trained using the training set and the performance is measured on the test set. As in the proposed CNN, a moving average filter with length 6 is used to remove the transient discontinuities.

2.5. Cluster-based Adaptive Sleep Staging (CLASS)

This algorithm is based on the relatively higher discontinuity in quiet sleep compared to non-quiet sleep and has been developed by Dereymaeker et al. [4]. Briefly, the pipeline of the algorithm is as follows. First, a pre-processing step consisting of a bandpass filter from 1 to 40 Hz and a notch filter at 50 Hz is performed. Artefact subspace reconstruction is then adopted to reject remaining artefacts, which could otherwise be wrongly detected as EEG discontinuities. After artefact removal, the cleaned EEG is downsampled with a factor three (sampling frequency = 83 Hz) to reduce the computational time during the subsequent steps. To deal with the nonstationarity of the EEG signal, the recordings are adaptively segmented resulting in quasi-stationary segments with variable length. Features extracted from these segments are then used to cluster segments with similar characteristics. A cluster time profile is obtained by representing each sample by its cluster label. Once the cluster time profile is extracted for each EEG channel, some processing steps are performed to obtain a single smooth envelope which can be thresholded to detect quiet sleep segments. This threshold is calculated as the mean of the envelope plus a factor multiplied with the standard deviation of the signal envelope. In the final stage of the algorithm, quiet sleep detections shorter than three minutes are removed [41].

2.6. Classification performance

In order to assess the performance of the sleep stage classifier, multiple performance measures have been computed. Sensitivity (Sen) is the probability of a detecting a quiet sleep segment as quiet sleep, which is computed as:

$$Sen = \frac{TP}{TP + FN}. \tag{3}$$

The specificity (Spe) is the likelihood of correctly classifying a non-quiet sleep segment as non-quiet sleep, that is:

$$Spe = \frac{TN}{TN + FP}. \tag{4}$$

TP, true positive, is the number of correctly identified quiet sleep epochs. TN, true negative, is the number of correctly identified non-quiet sleep epochs. FN, false negative, represents the number of quiet sleep segments that are classified as non-quiet sleep. FP, false positive, is the number of wrongly classified non-quiet sleep segments.

Cohen's kappa coefficient is a statistical measure of the inter-rater agreement, while taking into account the agreement occurring by chance. Because it corrects for the agreement expected by chance, it is a more conservative measure than accuracy and more suitable for imbalanced datasets. Cohen's kappa coefficient κ is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{5}$$

where p_o represents the observed agreement and p_e denotes the expected agreement.

All of the above-mentioned evaluation metrics are measured for each recording of the test set, and the median and interquartile range are reported. The computation of these measures require a fixed decision threshold. In this study, the optimal threshold was defined so that the sensitivity equals the specificity on the training set. Since the threshold used in the CLASS algorithm is computed as threshold = mean(y) + factor × std(y) where y denotes the smooth envelope, the factor instead of the complete threshold is optimized on the training set.

In addition, the receiver operating characteristic (ROC) curve is computed for each recording in the test set. This allows to analyze the effectiveness of the classification without defining a fixed threshold. The mean and median ROC curve over all recordings are then obtained by taking the mean and median of the true positive rate and the false positive rate at a specific threshold. These mean and median ROC curves are plotted and the area under the curve (AUC) is computed.

2.7. Error correlation

In order to investigate the error correlation between the three methods, the percentage of test segments correctly classified by all three algorithms, by only two or one of the algorithms, or by none of the algorithms are computed. In addition, Cohen's kappa is computed among each of the sleep classification algorithms and the clinical sleep labels in order to investigate the agreement between the different algorithms.

2.8. Computational time

It is generally known that deep neural networks are highly complex and computationally expensive to train. However, when using a sleep stage classifier in clinical practice, it is not the training time, but the time required to classify a new incoming EEG segment that is of interest.

In order to compare the computational evaluation time of the three algorithms, sleep stage classification of the complete test set in blocks of 2 h of EEG is performed with each of the algorithms. The three algorithms were run chronologically and the mean and standard deviation of the computational time are reported to avoid variable CPU loading. Since the complete test set is classified, more than 100 iterations are performed. This experiment was conducted on a workstation with an Intel(R) Core(TM) i7 3.6 GHz processor and 16 GB RAM, implemented in MATLAB R2016a software (The MathWorks, Natick, MA, USA).

3. Results

3.1. Feature evolution during sleep-wake cycling

To illustrate how each of the 10 features extracted by the convolutional neural network behave during each of the sleep states, boxplots of the features during sleep-wake

Table 2. The classification performance for the proposed CNN, the CLASS algorithm and the feature-based approach without (NP) and with (PP) post-processing respectively. The area under the mean ROC curve (AUC) and median(IQR) of the sensitivity, specificity and Cohen's kappa are set out.

	AUC	Sensitivity	Specificity	Kappa
CNN	0.92	0.88(0.26)	0.93(0.13)	0.74(0.17)
CLASS	0.92	0.72(0.19)	0.97(0.06)	0.75(0.19)
Feature-based NP	0.83	0.74(0.21)	0.87(0.13)	0.59(0.25)
Feature-based PP	0.93	0.83(0.28)	0.97(0.07)	0.77(0.23)

cycling are shown in Figure 3. The top of Figure 3 illustrates the amplitude integrated encephalography (aEEG) derived from the bicentral channels C3/C4 from a test recording at 31 weeks and 3 days postmenstrual age. aEEG is commonly used in the NICU to monitor the functional brain integrity and is a suitable tool to assess sleep-wake cycling of the infant [42]. More discontinuous activity during quiet sleep is recognized by a widening of the aEEG trace, while the more continuous acitivity during active sleep or wakefulness is characterized by a narrow trace [42, 43]. The aEEG of the quiet sleep segment selected in Figure 3 has a clearly wider bandwidth compared to the non-quiet sleep segments before and after. The top row of boxplots shows features 1, 3, 5, 7, 9 and 10, which are reduced during quiet sleep (the middle boxplot) compared to non-quiet sleep (left and right boxplot). The bottom row provides the provides the boxplots of features 2, 4, 6 and 8, which increase during quiet sleep. From the boxplots in Figure 3, it can be seen that most of the features are well discriminating quiet sleep from non-quiet sleep.

3.2. Classification performance

The classification performance of the proposed CNN is illustrated in Figure 4, which presents the ROC curves for each of the test recordings separately, and the mean and median ROC curve. The area under the mean and median ROC curve are equal to 92% and 98% respectively. The histogram of the training data segments and the AUC for each recording of the test set as function of postmenstrual age are plotted in Figure 5. It can be seen from the histogram that training data is available at all ages. From the AUC in function of PMA, it is apparent that most of the test recordings are measured between 30 and 38 weeks of PMA. Moreover, the figure shows that the algorithm performs well over a wide range of PMA. It is notable that in Figure 4 as well as in Figure 5 two outliers with a low AUC ($\leqslant 0.65$) can be observed. Table 2 compares the AUC, sensitivity, specificity and Cohen's kappa across the 3 methods. It can be seen from the table that the proposed CNN has similar performance compared to the CLASS and feature-based algorithm. In addition, it is clear that the post-processing step significantly improves the performance of the sleep stage classifier.

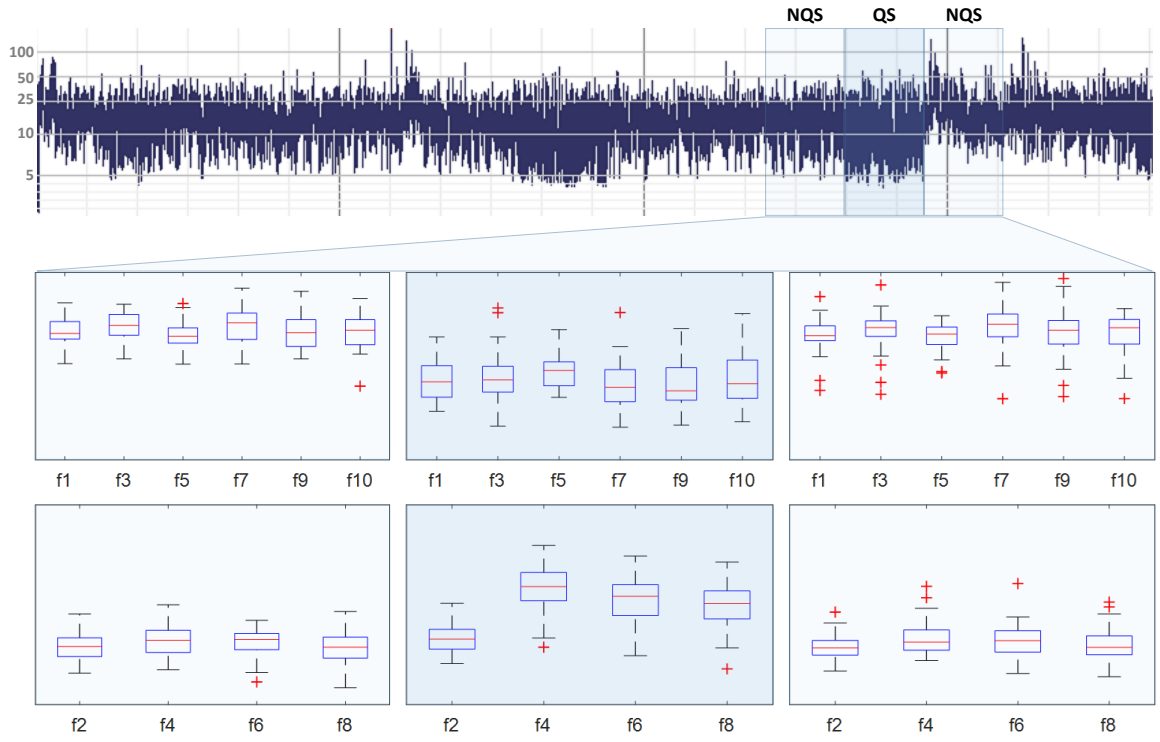


Figure 3. The aEEG trace derived from the bicentral channels C3 and C4 from a recording of the test set at 31 weeks and 3 days PMA is shown on top. The selected quiet sleep (QS) segment and preceding and subsequent non-quiet sleep (NQS) segments are indicated on the aEEG. The boxplots show the features during each of these selected segments. The top half of the boxplots show the six features that reduce during quiet sleep (f1, f3, f5, f7, f9, f10), the bottom half shows the four features that increase during quiet sleep (f2, f4, f6, f8). Most of the features are significantly different for the two sleep stages.

3.3. Error correlation

Figure 6(a) illustrates the percentage of segments correctly labelled by all the algorithms, by only two or one of the algorithms, or missclassified by all methods. The complete bars represent the complete test set, the dark grey part corresponds to the segments of the two outliers in Figure 5, while the light grey part of the bars correspond to the remaining segments. From this bar graph, it is clear that most of the segments (nearly 80%) are correctly classified by all algorithms. Approximately 14% of the segments are correctly identified by two of the algorithms, and only 8% of the segments were only correctly classified by one of the described methods. Less than 3% of the segments were wrongly classified by all the algorithms. However, a considerable amount of the segments that were misclassified by all algorithms are part of the two recordings with a low AUC. More specifically, from all segments of the outliers, almost 25% is not identified by any of the algorithms. This illustrates the high agreement among the three algorithms, which is

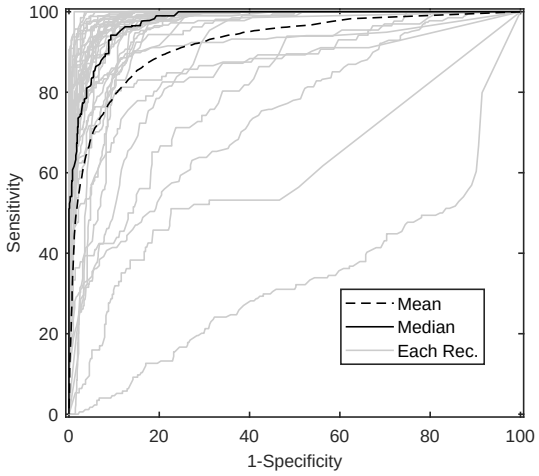


Figure 4. The ROC curves for the proposed CNN sleep stage classifier. The light grey ROC curves show the performance of the classifier for each recording of the test set. The black dashed and full line represent the mean ROC and median ROC curve respectively.

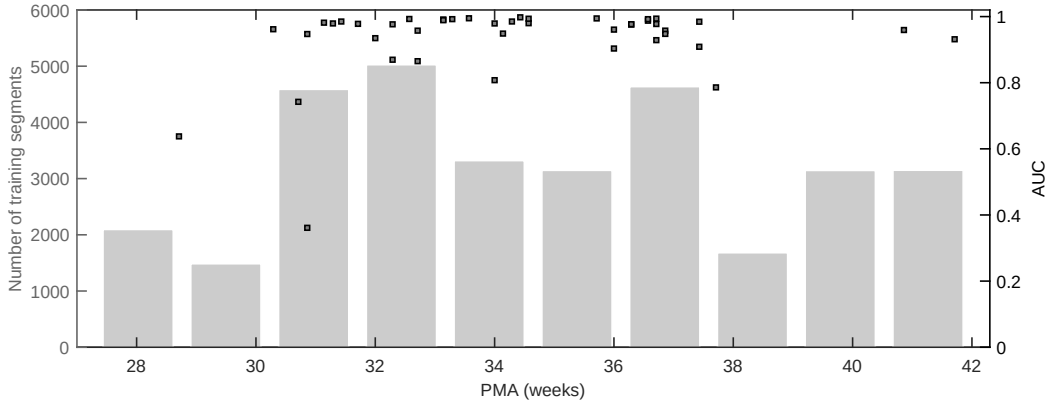


Figure 5. The histogram of the training data segments is displayed in light grey (left y axis). The squares show the AUC for each recording from the test set (right y axis).

also confirmed by the Kappa values shown in Figure 6(b).

3.4. Computational Time

The bar graph in Figure 7 illustrates the mean evaluation time for 2 h EEG segments. The dark blue box represents the average total time for loading, pre-processing and classifying 2 h EEG data. The light blue box shows the time required for classification. In case of CLASS, the classification consists of performing CLASS, starting from the artefact subspace reconstruction up to defining the quiet sleep periods. For the feature-based approach, the computational time for feature extraction, classification using the SVM and the post-processing are merged in the classification time. Finally, the time for classification and post-processing are combined and marked as classification time for

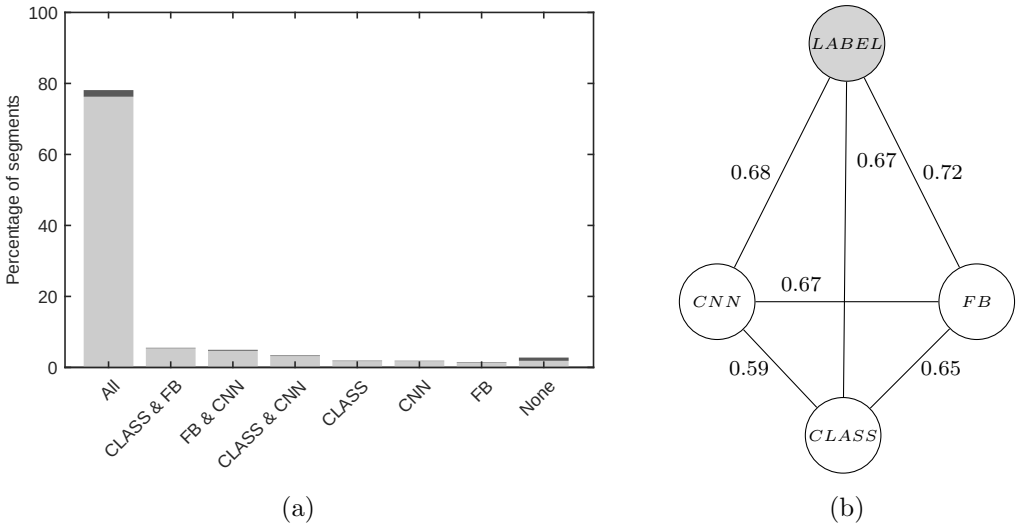


Figure 6. The error correlation between the three sleep stage classifiers. (a) The bar graph shows the percentage of EEG segments correctly classified by all, two, only one or none of the algorithms. The dark grey part of the bars correspond to the segments of the two outliers, while the light grey part corresponds to the remaining test segments. (b) This graph represents the agreement among the three algorithms (CNN, FB, CLASS) and the clinical labels (LABEL) computed using Cohen's kappa coefficient.

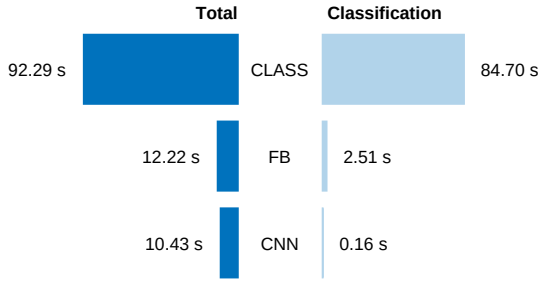


Figure 7. The average computational time for 2 h multichannel EEG segments for each of the three algorithms. The total computational time (left) and the time for classification (right) are shown.

the CNN.

From this chart, it can be seen that overall the CNN (10 ± 1.0 s) is 9 times faster than CLASS (92 ± 3.8 s) and slightly faster than the feature-based (12 ± 1.3 s) approach. The most remarkable difference is the classification time, which is significantly lower for the CNN (0.16 ± 0.04 s) compared to CLASS (85 ± 3.74 s) and the feature-based classification (2.5 ± 0.28 s).

4. Discussion

The aim of the current study was to implement a convolutional neural network for neonatal sleep stage classification. By adopting convolutional neural networks which are able to learn relevant features automatically from the raw EEG data, the difficult and time-consuming process of selecting the proper feature set, which requires domain knowledge from experts, can be avoided. Moreover, the CNN optimizes the feature extraction and classification simultaneously. This novel data driven approach successfully identified quiet sleep in preterm infants with an average and median AUC of 92% and 98% respectively.

While this performance is comparable to the CLASS algorithm developed by Dereymaeker et al. [4], the CNN has some advantages compared to the cluster time profiles approach. First of all, except for bandpass filtering the proposed method does not use any advanced artefact removal, whereas the CLASS algorithm uses artefact subspace reconstruction as additional artefact rejection. Secondly, the proposed CNN can be retrained if additional data collection is performed while this is not straightforward for the CLASS algorithm. As sleep patterns will be better represented in a larger training database, the classification performance will improve. At last, it must be pointed out that the threshold used in the CLASS algorithm is based on the mean and the standard deviation of the signal envelope. Therefore, the complete or at least a representative part of the recording should be available before the classification can be performed. In contrast to CLASS, the proposed CNN uses a fixed threshold which can be decided based on the training set and does not need to be adjusted to the new recording which facilitates real time sleep scoring. Although the proposed CNN uses only 10 features as input for the fully connected classification stage, it reaches the same performance as the feature-based approach which is based on 72 features [10]. To conclude, it can be seen in Figure 7 that the presented CNN is considerably faster compared to the current state-of-the-art.

One of the key challenges of neonatal sleep stage classification is to deal with the fast alterations in the EEG patterns during sleep ontogenesis. It is apparent from Figure 5 that the network is able to train the age specific EEG characteristics, resulting in a good performance over a wide PMA range (30 - 38 weeks PMA). Due to the small number of recordings below 30 weeks and beyond 38 weeks PMA, we can not make any statements about the performance in those age ranges. Both in Figure 4 and 5, two outliers can be observed. These recordings also had a low AUC in the CLASS and feature-based approach. According to an expert clinician, the outlier at 28 weeks and 5 days (AUC equal to 65%) is due to poor EEG quality probably as a result of electrode impedance problems, while the low performance of the recording at 30 weeks and 6 days (AUC equal to 36%) is caused by an intravenous infusion motor artefact.

In this study, the overall AUC is computed as the area under the mean or median ROC curve. As explained above, this mean ROC curve is obtained by averaging the sensitivity and specificity at an incremental threshold between zero and one, which is

called threshold averaging [44]. It is important to note that this procedure of first computing the average ROC and then the AUC is not the same as computing the AUC for each recording separately and then taking the average. Computing the mean of individual AUCs is called vertical averaging, which corresponds to taking the average of the true positive rates at a fixed false positive rate [44]. In this application we opted for threshold averaging rather than vertical averaging, because implementation of the algorithm in a brain monitor for use in clinical practice requires the choice of one fixed threshold.

The present study has relevance for clinical practice. First of all, the proposed CNN has a low computational time, hence it is suitable for real time sleep stage monitoring. Real time sleep scoring is important in the NICU as it can aid in making the neonatal care more patient-driven and avoid disruption of the infant's sleep. In addition to implementation of the algorithm in a brain function monitor to promote neonatal sleep in the NICU, this algorithm is of great interest in maturation studies as we have to account for the cyclic nature of EEG during brain maturation analysis. Furthermore, the sleep EEG undergoes fast maturational changes during early brain development.

The main limitation lies in the fact that the training phase is computationally expensive and requires a large database. However, training time is secondary since the network only has to be trained once. After training, the recall computational time is the important parameter, which is much smaller for CNN compared to the existing algorithms. A weakness of this study was the paucity of EEG recordings below 30 weeks and beyond 38 weeks postmenstrual age. Another drawback of CNNs is the lack of interpretability of the features. Further work needs to be done to establish which features are extracted by the network.

The results of the current study suggest that convolutional neural networks are a promising approach for sleep stage classification in preterm infants. However, several steps can be explored to further improve performance the sleep stage classification. A number of recommendations for future research are given. First of all, as mentioned before, the current architecture of the network has been defined based on trial and error. As a result, there is abundant room for further progress in designing the optimal network, especially with the aid of the high-speed graphics processing unit (GPU). In addition, the proposed method uses a post-processing step to take into account that the sleep state can not change instantly. In future investigations, it might be possible to incorporate this in the network by including information of the preceding and subsequent segments by using another deep learning algorithm such as Long Short Term Memory networks (LSTM). Moreover, clinicians do not rely only on EEG characteristics, but assess other non-cerebral criteria as well during visual sleep scoring [4, 45]. Behavioural correlates that are essential to distinguish quiet sleep and non-quiet sleep are body movements, eye movements and cardiorespiratory regularity. Therefore, we expect that the performance can be improved by including additional modalities such as EMG, respiration and EOG. Our results indicate that the proposed network can reliably classify sleep in preterm infants with a PMA from 30 to 38 weeks. Future studies should aim to replicate the

results for a larger age range, by retraining the network using data from an older, term neonates population. At last, interpretation of the features extracted by the network is an important issue for future research.

5. Conclusion

In this study, we designed and implemented a deep convolutional neural network which automatically extracts optimal features to discriminate quiet sleep from non-quiet sleep in preterm infants. The proposed network achieved the state-of-the-art performance over a wide range of PMA without using domain-specific knowledge. Furthermore, the presented sleep stage classifier has a low computational time, which makes the CNN based sleep stage classification suitable for real time sleep scoring in the NICU.

Acknowledgments

Research supported by Bijzonder Onderzoeksfonds KU Leuven (BOF): The effect of perinatal stress on the later outcome in preterm babies #: C24/15/036. imec funds 2017. European Research Council: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007–2013)/ERC Advanced Grant: BIOTENSORS(n°339804). 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 Dorado is a postdoctoral fellow at Fonds voor Wetenschappelijk Onderzoek-Vlaanderen (FWO), supported by Flemish government. Mario Lavanga is a SB PhD fellow at Fonds voor Wetenschappelijk Onderzoek-Vlaanderen (FWO), supported by Flemish government.

References

[1] Chiara Nosarti, Kie Woo Nam, Muriel Walshe, Robin M. Murray, Marion Cuddy, Larry Rifkin, and Matthew P.G. Allin. Preterm birth and structural brain alterations in early adulthood. *NeuroImage: Clinical*, 6:180–191, 2014.

[2] Rita H. Pickler, Jacqueline M. McGrath, Barbara A. Reyna, Nancy McCain, Mary Lewis, Sharon Cone, Paul Wetzel, and Al Best. A Model of Neurodevelopmental Risk and Protection for Preterm Infants. *The Journal of Perinatal & Neonatal Nursing*, 24(4):356–365, 2010.

[3] Mark S. Scher. Ontogeny of EEG-sleep from neonatal through infancy periods. *Sleep Medicine*, 9(6):615–636, 2008.

[4] Anneleen Dereymaeker, Kirubin Pillay, Jan Vervisch, Maarten De Vos, Sabine Van Huffel, Katrien Jansen, and Gunnar Naulaers. Review of sleep-EEG in preterm and term neonates. *Early Human Development*, 2017.

[5] Kimberly A Allen. Promoting and protecting infant sleep. *Advances in neonatal care*, 12(5):288–291, 2012.

[6] Agnes van den Hoogen, Charlotte J. Teunis, Rene A. Shellhaas, Sigrid Pillen, Manon Benders, and Jeroen Dudink. How to improve sleep in a neonatal intensive care unit: A systematic review. *Early Human Development*, 2017.

[7] Ninah Koolen, Lisa Oberdorfer, Zsofia Rona, Vito Giordano, Tobias Werther, Katrin Klebermass-Schrehof, Nathan Stevenson, and Sampsa Vanhatalo. Automated classification of neonatal sleep states using EEG. *Clinical Neurophysiology*, 128:1100–1108, 2017.

[8] Kirubin Pillay, Anneleen Dereymaeker, Katrien Jansen, Gunnar Naulaers, Sabine Van Huffel, and Maarten De Vos. Automated EEG sleep staging in the term-age baby using a generative modelling approach. *Journal of Neural Engineering*, 15(3), 2018.

[9] Marco Carrozzi, Agostino Accardo, and Furio Bouquet. Analysis of sleep-stage characteristics in full-term newborns by means of spectral and fractal parameters. *Sleep*, 27(7):1384–1393, 2004.

[10] Alexandra Piryatinska, Gyorgy Terdik, Wojbor A. Woyczynski, Kenneth A. Loparo, Mark S. Scher, and Anatoly Zlotnik. Automated detection of neonate eeg sleep stages. *Computer Methods and Programs in Biomedicine*, 95(1):31 – 46, 2009.

[11] Kirsi Palmu, Turkka Kirjavainen, Susanna Stjerna, Tommi Salokivi, and Sampsa Vanhatalo. Sleep wake cycling in early preterm infants: Comparison of polysomnographic recordings with a novel EEG-based index. *Clinical Neurophysiology*, 124(9), 2013.

[12] Nathan J Stevenson, Kirsi Palmu, Sverre Wikström, Lena Hellström-Westas, and Sampsa Vanhatalo. Measuring brain activity cycling (BAC) in long term EEG monitoring of preterm babies. *Physiological Measurement*, 35(7):1493–1508, 2014.

[13] V. Krajca, S. Petranek, K. Paul, M. Matousek, J. Mohylova, and L. Lhotska. Automatic Detection of Sleep Stages in Neonatal EEG Using the Structural Time Profiles. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 6014–6016. IEEE, 2005.

[14] Vladimír Krajča, Svojmil Petránek, Jitka Mohylová, Karel Paul, Václav Gerla, and Lenka Lhotská. Neonatal EEG Sleep Stages Modelling by Temporal Profiles. In *In: Moreno Díaz R., Pichler F., Quesada Arencibia A. (eds) Computer Aided Systems Theory EUROCAST 2007. EUROCAST 2007. Lecture Notes in Computer Science*, volume 4739, pages 195–201, Berlin, 2007.

[15] Anneleen Dereymaeker, Kirubin Pillay, Jan Vervisch, Sabine Van Huffel, Gunnar Naulaers, Katrien Jansen, and Maarten De Vos. An Automated Quiet Sleep Detection Approach in Preterm Infants as a Gateway to Assess Brain Maturation. *International Journal of Neural Systems*, 27(0):1750023, 2017.

[16] J S Barlow. Computer characterization of tracé alternant and REM sleep patterns in the neonatal EEG by adaptive segmentation–an exploratory study. *Electroencephalography and clinical neurophysiology*, 60(2):163–73, feb 1985.

[17] Luay Fraiwan, Khaldon Lweesy, Natheer Khasawneh, M Fraiwan, Heinrich Wenz, and Hartmut Dickhaus. Time frequency analysis for automated sleep stage identification in fullterm and preterm neonates. *Journal of Medical Systems*, 35(4):693–702, 2011.

[18] Mario Lavanga, Ofelie De Wel, Alexander Caicedo Dorado, Elisabeth Heremans, Katrien Jansen, Anneleen Dereymaeker, Gunnar Naulaers, and Sabine Van Huffel. Automatic quiet sleep detection based on multifractality in preterm neonates: effects of maturation. In *Proc. 39th Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, Seogwipo, South Korea, 2017.

[19] Ofelie De Wel, Mario Lavanga, Alexander Caicedo Dorado, Katrien Jansen, Anneleen Dereymaeker, Gunnar Naulaers, and Sabine Van Huffel. Complexity Analysis of Neonatal EEG Using Multiscale Entropy : Applications in Brain Maturation and Sleep Stage Classification. *Entropy*, 19(516), 2017.

[20] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[21] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, and Gerald Penn. Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4277–4280. IEEE, 2012.

[22] Dan C Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, and Jürgen Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In *IJCAI*

Proceedings-International Joint Conference on Artificial Intelligence, volume 22, page 1237. Barcelona, Spain, 2011.

[23] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1):98–113, 1997.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[26] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep Learning in Medical Image Analysis. *Annual review of biomedical engineering*, 19:221–248, jun 2017.

[27] U. Rajendra Acharya, Hamido Fujita, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Muhammad Adam. Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. *Information Sciences*, 415-416:190–198, 2017.

[28] U. Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Hojjat Adeli. Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Computers in Biology and Medicine*, 2017.

[29] A. H. Ansari, P. J. Cherian, A. Caicedo, G. Naulaers, M. De Vos, and S. Van Huffel. Neonatal seizure detection using deep convolutional neural networks. *Accepted for publication in International Journal of Neural Systems*, 2018.

[30] Orestis Tsinalis, Paul M. Matthews, Yike Guo, and Stefanos Zafeiriou. Automatic Sleep Stage Scoring with Single-Channel EEG Using Convolutional Neural Networks. *ArXiv*, oct 2016.

[31] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. DeepSleepNet: a Model for Automatic Sleep Stage Scoring based on Raw Single-Channel EEG. *IEEE transactions on neural systems and rehabilitation engineering*, mar 2017.

[32] Kaare Mikkelsen and Maarten De Vos. Personalizing deep learning models for automatic sleep staging. arXiv:1801.02645 [q-bio.NC].

[33] Stanislas Chambon, Mathieu Galtier, Pierrick Arnal, Gilles Wainrib, and Alexandre Gramfort. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *arXiv:1707.03321*, pages 1–12, 2017.

[34] Arnaud Sors, Stéphane Bonnet, Sébastien Mirek, Laurent Vercueil, and Jean François Payen. A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomedical Signal Processing and Control*, 42:107–114, 2018.

[35] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification. *arXiv:1805.06456v1*, 2018.

[36] Perumpillichira J Cherian, Renate M Swarte, and Gerhard H Visser. Technical standards for recording and interpretation of neonatal electroencephalogram in clinical practice. *Annals of Indian Academy of Neurology*, 12(1):58–70, jan 2009.

[37] Jake Bouvrie. Notes on convolutional neural networks. 2006.

[38] Thomas Serre, Aude Oliva, and Tomaso Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the national academy of sciences*, 104(15):6424–6429, 2007.

[39] Thomas Serre, Minjoon Kouh, Charles Cadieu, Ulf Knoblich, Gabriel Kreiman, and Tomaso Poggio. A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. Technical report, AI MEMO MIT, 2005.

[40] Andrea Vedaldi and Karel Lenc. Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 689–692. ACM, 2015.

[41] Irina Korotchikova, Nathan J Stevenson, Vicki Livingstone, C Anthony Ryan, and Geraldine B Boylan. Sleep-wake cycle of the healthy term newborn infant in the immediate postnatal period.

Clinical neurophysiology, 127(4):2095–2101, 2016.

[42] L S de Vries and L Hellström-Westas. Role of cerebral function monitoring in the newborn. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 90:F201–FF207, 2005.

[43] Daphna Yasova Barbeau and Michael D. Weiss. Sleep Disturbances in Newborns. *Children*, 4(10):90, 2017.

[44] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[45] Daniel L Picchietti, Stephan H Sheldon, and Conrad Iber. The Visual Scoring of Sleep and Arousal in Infants and Children. *Journal of Clinical Sleep Medicine*, 3(1):201–240, 2007.