A convolutional neural network outperforming state-of-the-art sleep staging algorithms for both preterm and term infants

Amir H. Ansari [1,2] *, Ofelie De Wel [1,2], Kirubin Pillay [3], Anneleen Dereymaeker [4], Katrien Jansen [4,5], Sabine Van Huffel [1,2], Gunnar Naulaers [4], Maarten De Vos [3]

Abstract— **Objective: To classify sleep states using electroencephalogram (EEG) that reliably works over a wide range of preterm ages, as well as term age. Methods: A convolutional neural network is developed to perform 2- and 4-class sleep classification in neonates. The network takes as input an 8-channel 30-second EEG segment and outputs the sleep state probabilities. Apart from simple downsampling of the input and smoothing of the output, the suggested network is an end-to-end algorithm that avoids the need for hand-crafted feature selection or complex pre/post processing steps. To train and test this method, 113 EEG recordings from 42 infants are used. Results: For quiet sleep detection (the 2-class problem), mean kappa between the network estimate and the ground truth annotated by EEG human experts is 0.76. The sensitivity and specificity are 90% and 88%, respectively. For 4-class classification, mean kappa is 0.64. The averaged sensitivity and specificity (1 vs. all) respectively equal 72% and 91%. The results outperform current state-of-the-art methods for which kappa ranges from 0.66 to 0.70 in preterm and from 0.51 to 0.61 in term infants, based on training and testing using the same database. Significance: The proposed method has the highest reported accuracy for EEG sleep state classification for both preterm and term age neonates.**

Index Terms— **Automated sleep stage classification; quiet sleep detection; deep learning; convolutional neural networks**

I. Introduction

Some newborn babies need special nursing and monitoring after delivery in the neonatal intensive care unit (NICU), e.g. due to asphyxia, respiratory and cardiac disorders, sepsis or infection, low birth weight, and prematurity [1]–[3]. Premature, or preterm neonates refer to infants who are born before 37 postmenstrual age (PMA), where PMA is the number of weeks elapsed since the first day of the last menstrual cycle of the mother till the time of recording. Generally, prematurity is one of the main causes of infant mortality, and affected infants comprise the majority of babies cared for in the NICU [4], [5]. While these vulnerable infants are in the NICU, bed-side neuromonitoring systems can provide valuable insights into their neurological development and brain maturation [6]–[8]. To this end, electroencephalogram (EEG) is used as a reliable and relatively easy way to non-invasively follow the neurological state of the neonate [9]. However, special expertise is needed to interpret the EEG recordings which is costly and not available around the clock in most centres. Therefore, automation of EEG analysis through computerised algorithms, artificial intelligence, and machine learning can potentially improve neurodevelopmental diagnostics and treatment [9].

Sleeping is the primary activity of newborns, particularly preterms, and has important roles in the development and maturation of cortical pathways, structural development, and optimal physical growth [10]. Initial evidence of sleep-wake cycling manifests in the EEG around 27 weeks PMA. However, clear sleep patterns predominantly appear after 31 weeks PMA. In preterms, the sleep period is initially divided into two states as active sleep (AS), also called rapid eye movement sleep (REM), and quiet sleep (QS), also referred as non-REM [9]. The EEG morphology of AS is very similar to wakefulness, as both show continuous traces. However, those can be discriminated based on eye and motor movements, which can be measured by polysomnographic signals, e.g. electromyogram (EMG) and electro-oculogram (EOG). On the other hand, QS exhibits discontinuous traces, consisting of burst cycles of high amplitude separated by inter-burst intervals (IBIs) with electrographic quiescence [8], [9], [11].

At the onset of term age, after 36 weeks PMA, each of AS and QS can be divided into two sub-states: 1) ASI with anterior

[1] Department of Electrical Engineering (ESAT), STADIUS, KU Leuven, Belgium

[2] imec, Leuven, Belgium

[3] Institute of Biomedical Engineering, Department of Engineering, University of Oxford, Oxford, UK

[4] Department of Development and Regeneration, University Hospitals Leuven, Neonatal Intensive Care Unit, KU Leuven, Leuven, Belgium

[5] Department of Development and Regeneration, University Hospitals Leuven, Child Neurology, KU Leuven, Leuven, Belgium

* correspondence e-mail: amirhossein.ansari@kuleuven.be

AHA, ODW, and SVH are supported by: Bijzonder Onderzoeksfonds KU Leuven (BOF): The effect of perinatal stress on the later outcome in preterm babies #: C24/15/036; Fonds voor Wetenschappelijk Onderzoek-Vlaanderen (FWO): Hercules Foundation (AKUL 043) 'Flanders BCI Lab - High-End, Modular EEG Equipment for Brain Computer Interfacing; imec funds 2017; EU: European Union's Seventh Framework Programme (FP7/2007-2013) The HIP Trial: #260777; EU H2020 MSCA-ITN-2018: 'INtegrating Functional Assessment measures for Neonatal Safeguard (INFANS)', funded by the European Commission under Grant Agreement #813483. AHA is furthermore supported by Internal Funds KU Leuven (PDM/18/143) and FWO postdoctoral fellowship. KP and MDV are supported by Wellcome Trust Centre, (Sleep, Circadian Rhythms & Neuroscience Institute, 098461/Z/12/Z).

delta activity, namely 'anterior dysrhythmia', 2) ASII or low voltage irregular (LVI), with low-amplitude and rapid theta as well as alpha activity, 3) QSI, or high voltage slow-wave (HVS) with high-amplitude occipital and central delta activity, and finally 4) QSII, or 'Tracé Alternant' (TA) with equal length of bursts and IBIs [8], [9]. Figure 1 illustrates two 20-second segments of EEG from a recording at 39 weeks PMA where Figure 1 (A) shows a TA with discontinuous burst/inter-burst pattern and Figure 1 (B) is an ASI segment with an anterior dysrhythmia. The unclear intervals between sleep stages are referred as transitionary sleep (TS), when multiple components of AS and QS, or the sub-states, manifest at the same time. This transition time shortens, QS periods lengthens, and the proportion of AS reduces whilst the infant is maturing [8], [11]. The age-dependency of neonatal sleep states and the evolution of their morphological patterns are among the challenges in developing an automated and robust.

To automate neonatal EEG sleep state classification in preterms and term infants, different characteristics of EEG have been quantified and considered, including EEG (dis)continuity [12], frequency content of EEG [13], proportional duration of bursts [14], and the frequency content of bursts [15]. Furthermore, in other studies, the extracted features from adaptively segmented EEGs are temporally analysed [16]–[18]. This time profile analysis, called cluster-based adaptive sleep staging (CLASS), has been improved for preterm QS detection in [19] and tested on term babies in [20]. In addition, Pillay et al. demonstrated the use of features in frequency and time domain, totalling 112 features, as part of a hidden Markov model (HMM) and Gaussian mixture model (GMM) for classification in term infants [20].

In all considered methods, in addition to different pre/post-processing steps, the features were engineered and selected by the human developers, which may not be optimal. As an alternative, deep neural networks provide a framework for 'end-to-end' learning, using the raw EEG directly as the input, without the need for hand-crafted feature extraction or complex pre/post-processing stages. Recently, a convolutional neural network (CNN) has been proposed in our group in [21] for preterm QS detection. However, due to its simple architecture focussed on preterm babies, it is not suitable for accurate term sleep staging.

In this paper, a newly designed CNN with a more complex and efficient structure is proposed that can not only detect QS in preterms more accurately (2-class classification), but can also detect QS in term babies. Moreover, this network is able to classify the four emerging stages of sleep in term neonates (4-class classification), which is important for monitoring maturation [8]. This results in the first end-to-end method based on deep learning that can classify sleep states in both preterm and term infants.

II. Material and Methods

A. Database

Data was obtained at the NICU of the University Hospitals of Leuven (UZ Leuven), Belgium and approved by the medical ethics committee of UZ Leuven. Informed parental consent was always obtained. The EEG recordings used 9 electrodes: Fp1, Fp2, C3, C4, Cz, T3, T4, O1, and O2. Among them, Cz is used as reference (totalling 8 mono-polar channels), based on the international 10-20 system [9]. The data was recorded at 250 Hz, using the BrainRT EEG recording system (OSG BVBA Rumst, Belgium). All newborns had normal neurodevelopmental outcome at 9 and 24 months of age, according to the criteria set out in [20], [21]. Two datasets were used in this study:

1) Preterm dataset: this dataset consists of 97 multichannel EEG recordings from 26 prematurely born infants recorded between 2012 and 2014. All babies were born before 32 weeks PMA and each of them had at least two recordings. The quiet sleep segments were identified by two EEG experts (AD and KJ) upon consensus. The non-quiet sleep segments include active sleep, wakefulness, and indeterminate segments (IS). This dataset was previously used in [19], [21] as well.

2) Term dataset: this dataset includes 16 recordings from 16 infants recorded at term age (8 born preterm). In this dataset, one expert EEG reader (AD) annotated the sleep (sub-)states as AS (ASI and ASII), wakefulness, QS (TA and HVS), IS, as well as artefact segments. The second expert (KJ) then reviewed these annotations and 'dubious' segments were defined as those where the experts disagreed. This annotating process was performed using neural (EEG) and behavioural (EMG, ECG, and respiration) signals. Since both ASII and wakefulness include LVI patterns in EEG, which are not separable using EEG alone [9], those are merged into LVI state. Dubious, IS, and artefact segments were excluded from further analysis. Therefore, the final labels are: ASI, HVS, TA, and LVI. Similar annotations for this dataset were previously used in [20].

B. Data preparation for training, validating and testing

In the preterm dataset for 2-class classification, the training and testing subsets were defined based on a fixed split that was previously used in [19], [21]. This results in a fair comparison when the proposed algorithm is compared with the state-of-the-art methods considered in those studies. All EEGs are bandpass filtered between 1 and 15 Hz and afterwards downsampled to 30 Hz. Then, they are segmented into 30-second epochs with no overlap. In total, this leads to approximately 32K and 26K EEG segments for training and testing the QS detection, respectively, in this preterm dataset. In order to validate the proposed algorithm during the design process, 'leave-one-subject-out' (LOSO) cross-validation is applied within the training dataset. The final method is then tested on the test dataset and the performance metrics are measured.

In the term dataset, the same filtering, downsampling, and segmenting techniques are used. For 2-class classification, the model trained on the preterm data is used with no extra training, while for the 4-class classification a new training round is

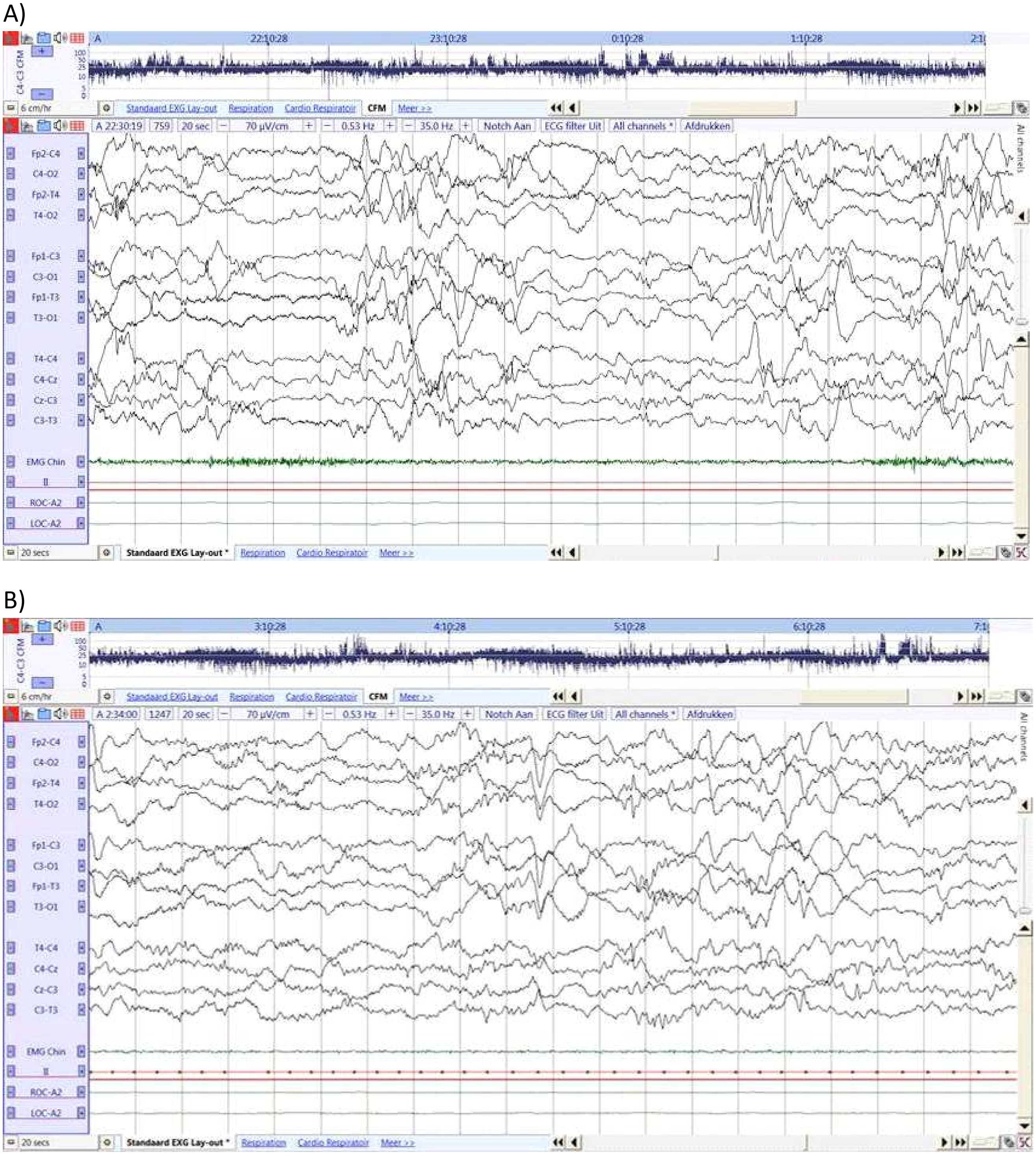


Figure 1: Two examples of a 20-second EEG segment in bipolar montage. Both segments are part of a recording at 39 weeks PMA from a term-born infant (GA: 38 weeks). The top one (A) was scored as 'Tracé alternant' having clear discontinuous burst/inter-burst patterns. The bottom one (B) was labeled as ASI and is characterized by anterior dysrhythmia.

needed. In this case, due to the need for a larger training dataset, instead of a fixed split between training and testing, LOSO is used across all data and the averaged performance is reported. This technique was similarly applied for developing and training the compared HMM and GMM methods in [20]. In each iteration of the LOSO, 20% of the training data are used as validation to avoid overtraining the network ('early stopping').

C. The state-of-the-art reference methods

In this study, 5 state-of-the-art algorithms that were previously developed and validated based on the same datasets are compared with the proposed algorithm. These methods are summarized below. It is important to note that in developing, training, and testing the following state-of-the-art algorithms (for both preterm and term problems), the same datasets, as described in database subsection, were used. This helps to have a fair comparison between these methods and the one proposed in this paper.

Cluster-based adaptive sleep staging (CLASS): This algorithm was proposed in [19] and was developed using the preterm dataset. In this method, bandpass filtered EEG is first cleaned of high-power muscle artefacts using the artefact subspace reconstruction technique. The cleaned EEG is then adaptively split into quasi-stationary segments and 9 frequency and time-domain features are extracted from each segment. All segments are clustered into 12 groups using k-means clustering forming 'cluster-time profiles' depicting the changes in the cluster labels with time. These cluster-time profiles were then smoothed over time and averaged across EEG channels before thresholding to define the QS segments where the profile is above the threshold. In general, this method detects EEG segments assuming that large changes represent greater EEG discontinuity [19]. Although this method has a suitable performance when applied to preterm babies [19], its performance was limited for term QS detection [20], as is further reported in this paper as well.

Feature-based QS detection (FBD): This algorithm is described in [21] for preterm 2-class classification. First, EEG is split into 30-second segments with no overlap. Then, 9 spectral features proposed in [13] are extracted from each EEG channel. The features from all 8 channels (totalling 72) are fed into a support vector machine (SVM) and the probabilistic output of the SVM for each epoch is smoothed by an averaging filter across time and thresholded to define the QS segments [21].

Gaussian mixture model (GMM) and Hidden Markov model (HMM): In these methods, which are proposed in [20] for sleep staging in term infants, the EEG is first filtered and split into 30-second segments with 5 seconds overlap. Then, 112 features are extracted from each segment of each EEG channel and the median is taken across channels. Next, the best features are selected using minimum redundancy maximum relevance. Finally, the selected features are fed into both a GMM and an HMM to classify the four sleep states. This paper further suggests a patient-wise rescaling of the features before feeding them into the classifier in order to improve performance. The rescaling means correcting the inter-recording variabilities by standardizing the features with the means and standard deviations of the features from the corresponding recording. Using this rescaling approach needs the features of the whole EEG recording to be available before the start of sleep staging. Both the GMM and HMM with/without rescaling, are reported and compared with the proposed method.

Previously developed CNN (CNNpre): This method was developed in [21] for preterm 2-class classification. In this method, the EEG was downsampled to 30Hz and divided into 30-second segments. The multichannel EEG segments are then fed into a CNN with 17 layers and 3027 trainable parameters. The output is smoothed with an averaging filter. In the current paper, the architecture of this network is changed, the layers are empowered with batch-normalization and drop-out layers, the complexity is increased, a better optimizer algorithm is applied, and the output and training are changed to be able to detect 2 and 4 sleep states in preterms and terms, as explained in the next section.

D. Proposed method

Convolutional neural networks: Convolutional neural networks (CNNs) are a type of artificial neural networks (ANN), consisting of alternating stacked convolutional (Conv), nonlinear, and pooling layers. A Conv layer is an extended version of a finite impulse response (FIR) filter bank as ($\boldsymbol{O_k} = f_k * I$) where I is the 3-dimensional input tensor, f_k is the k^{th} filter of the filter bank, O_k is the output of the corresponding k^{th} filter, and '$*$' denotes the convolution operation that applies on the first and second modes of the inputs. Since the hidden layers of a multilayer ANN must be nonlinear, a nonlinear operator is employed usually after each Conv layer. In the literature, the most common nonlinear unit used in CNN structures is the rectified linear unit (ReLU) (or its variations), which is the common half-wave rectifier that keeps the positive values unchanged and replaces the negative values by zero ($\max(0, x)$). In order to reduce the amount of trainable parameters and, therefore, control overfitting, a pooling layer is used after successive Conv layers. A pooling layer is an aggregator downsampling the output volume of its previous layer. The most common pooling operators are 'maxpool' and 'avgpool' which downsample the data by respectively taking the maximum and average of the data samples in each window. In classification tasks, these layers are usually followed by a couple of dense layers, which is a classic multilayer perceptron (MLP) with fully connected layers. In this case, the CNN automatically extracts features and the MLP, normally with one or two hidden layers, performs the classification. In addition to these main layers, other types of layers may be used depending on the problems, including: batch-normalization (to standardize the extracted feature maps) [22], drop-out (to increase the generalization of the dense layers) [23], and softmax (to exponentially normalize the network outputs and represent them as probabilities corresponding to the target classes), etc. (See [24], [25] for more details.) Although the general CNN structure and the main functionalities originated in image processing studies, a growing body of literature has recently investigated different CNN architectures for EEG analysis, e.g. for seizure detection [26]–[28], evoked response potential classification [29], [30], sleep analysis [21], [31]–[35], decoding task-related EEG information [36], and EEG-based auditory attention detection [37].

Proposed architecture: In this study, a new, efficient, and robust architecture with 21K parameters (about seven times

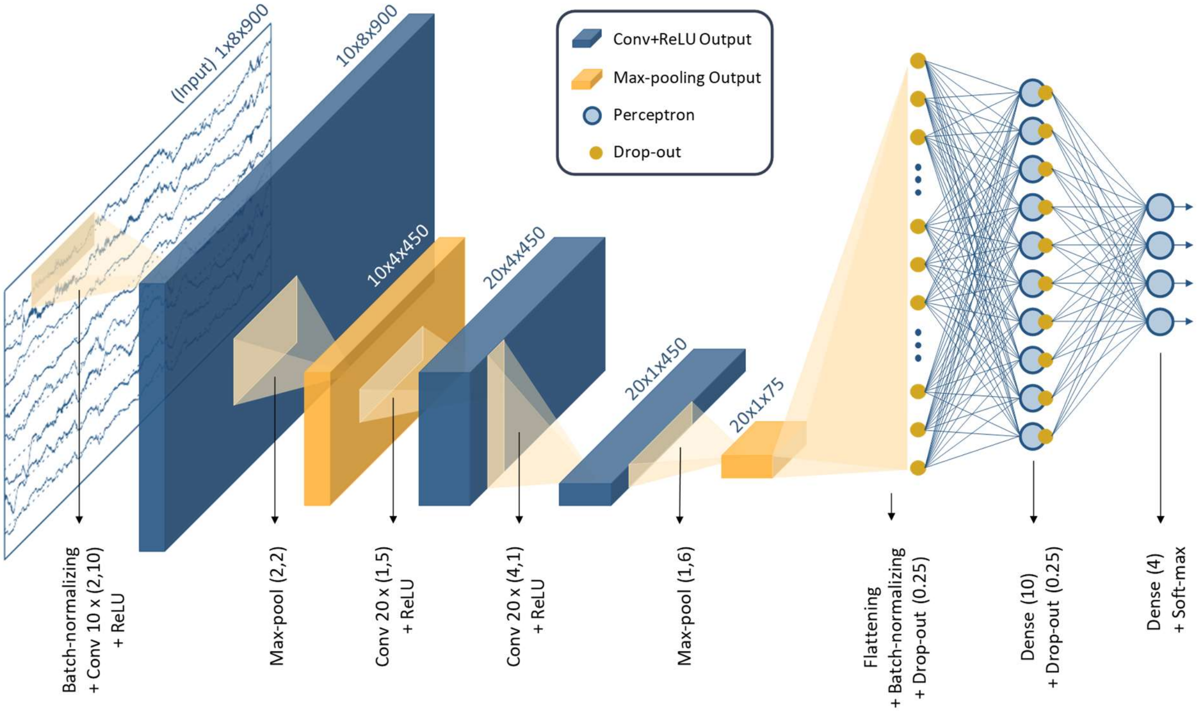


Figure 2: Structure of the proposed CNN. The input is an 8-channel, 30-second EEG segment and the output is 4 probabilities corresponding to the target classes.

more complexity than the previous network) with a better aggregation on EEG channels is proposed for preterm (2-class) and term (2 and 4-class) sleep staging. In this network with a large number of parameters, multiple batch-normalization and drop-out layers, as described below, are used to avoid overfitting.

Figure 2 summarizes the proposed CNN structure. The input is an 8-channel 30-second EEG segment which is initially downsampled to 30 Hz. The first Conv layer, which follows a batch-normalizer, has 10 kernels of (2×10), where the 1st mode operates on the channels (spatial) and the 2nd mode operates on data-points (temporal). In the spatial mode, involving 2 channels empowers the network to generate bipolar EEG channels to emphasize on contralateral neural activities if needed, similarly to how human expert EEG interpreters perform in some EEG analyses. In the temporal mode, 10 time-samples corresponding to one-third of a second seem short enough to extract local features, as the first filtering step, but not very short to extract natural EEG noise-like patterns. In this layer, zero-padding was used to keep the size of input and output data the same. This layer is followed by a ReLU and a max pooling layer with size (2×2), respectively. As a result, the spatial and temporal dimensions reduce to 4 channels and 450 data-points in 10 feature maps. Then, a single channel Conv layer with 20 kernels of size (1×5) is employed. Similarly, zero-padding preserves the size of volumes and the ReLU adds further nonlinearity. This precedes the third Conv layer with 20 kernels of (4×1) with no zero-padding, which combines all 4 bipolar channels into a single channel. Then, another ReLU and max pooling with size (1×6) are utilized, which produces 20 feature maps of 75 time samples. A batch-normalizer standardizes the extracted features. Next, all cells of the feature maps (totalling 1500 cells) are rearranged in a single vector (flattening) and are fed into a two-layer MLP with two drop-out layers (each 25% dropping chance). The number of output neurons is 2 for 2-class classification in preterm and term datasets and 4 for 4-class classification in the term dataset.

At the end of this process, a simple averaging filter is applied on the probabilistic outputs of the CNN from chronological EEG input segments. Since the sleep states do not change rapidly, this smoothing process increases the robustness of the outputs and reduces the effect of short artefacts. The length of this smoothing filter will be discussed in the next sections.

Channel dropping analysis: In practice, different number of EEG channels may be recorded and used for sleep stage classification. In order to evaluate the performance of the proposed method for different numbers of EEG channels, two analyses are performed: I) Dropping and retraining: in this analysis, the number of EEG channels is reduced from 8 to 4 (C3, C4, O1 and O2), 2 (C3 and C4), and 1 (C3 – C4). These channels are selected because these are commonly used EEG electrodes in cerebral function monitors (CFM^{TM}). Then, the CNN is retrained on the training data and the test performance is benchmarked against the CNN with all 8 channels. II) Dropping and testing: The second analysis is to simulate the case when some EEG channels are detached during a test while the applied method is unchanged. To this end, all possible channel dropping combinations (2^8-1) are made in the testing dataset and the performance of the CNN (without retraining) on these combinations is measured. In this analysis, in order to keep the dimension of the CNN input unchanged when different numbers of channels are eliminated, the rows corresponding to the detached channels are replaced by the average of the remaining ones. The performance of these two analyses are reported in the next section.

Performance metrics: In preterm QS detection (2-class problem), sensitivity, specificity, the area under the 'receiver operating characteristic' curve (AUC), and Cohen's kappa are used. These metrics are reported as mean ± standard deviation (SD) calculated over the recordings.

In 4-class classification, sensitivity, specificity, and AUC, are calculated as one class versus the others (one vs. all) and the average values over the classes are computed (the 'macro-average') [20]. Kappa is calculated directly since it is a multi-class metric. Similarly to the QS detection, the mean ± standard deviation of these metrics calculated over all recordings are reported. Furthermore, in order to test the improvement made by the proposed method per recording, a bootstrap test that was introduced in [20] is applied. To this end, the proportion of the bootstrapped recordings in the test database, for which the corresponding algorithm reaches higher performance than the proposed CNN is calculated and reported as p_{boot}. Therefore, p_{boot} values smaller than 50% indicate the success of the proposed CNN against the corresponding models in the majority of recordings, and vice versa.

In addition, the confusion and confidence matrices are reported. For classification of a segment, the confidence of the CNN is defined as the difference between the probabilities of the classes with the two highest values. For instance, if the output of the softmax is {0.1, 0.4, 0.5, 0} for an arbitrary segment, the confidence equals 0.1 (= 0.5 - 0.4). Thus, the confidence equals 0 (minimum) when the two top most classes are equal regardless of their absolute values, e.g. {0.1, 0.3, 0.3, 0.2}, and equals 1 (maximum) if only one class is at 1 probability and the remaining classes are 0, e.g. {0, 1, 0, 0}. In other words, the confidence value defines the margin in which the probability of the winning class can reduce while the classification output remains unchanged. This value is measured per segment and subsequently the overall confidence matrix is formed taking into account the class labels and ground truth, as with the conventional confusion matrix.

III. Results

A. Two-class classification

Preterm recordings: The test performance of the proposed and state-of-the-art methods for QS detection (2-class problem) in preterm babies are listed in Table I. Kappa, sensitivity, specificity, AUC, and p_{boot} are calculated over the test recordings.

Term recordings: Table II lists the test performance of the proposed method compared to the state-of-the-art for the term recordings in the QS detection problem. In this table, HMM$_R$ and GMM$_R$ correspond to the HMM/GMM with patient-wise rescaling.

B. Four-class classification

Table III shows the overall test performance of the proposed method and the state-of-the-art methods with/without rescaling in the 4-class classification problem in the term dataset. Since sensitivity and specificity are only measurable in 2-class problems, the macro-average is calculated per recording and the mean and SD over the recordings are then reported.

The confusion matrix of the proposed method is shown in Figure 3 (A). Each cell includes the normalized value as a percentage and the total number of classified segments in parenthesis. The green shaded cells along the main diagonal show the correctly classified segments and the red shaded cells indicate the groups with highest number of falsely classified segments. The confidence matrix is shown in Figure 3 (B).

Figure 4 illustrates the hypnogram of two term recordings where the CNN performs poorly (A: kappa = 0.36) and almost perfectly (B: kappa = 0.87). In both cases, the top row shows the clinician's labels and the bottom displays the algorithm's output. The class labels are on the vertical axes and the horizontal axes represents the time in hours.

Table I
The classification performance of the proposed methods compared to the state-of-the-art for the QS detection in the preterm recordings

Model	Mean Kappa (SD)	Mean % Sens (SD)	Mean % Spec (SD)	AUC %	p_{boot}
CLASS [19]	0.66 (0.24)	69 (20)	95 (06)	92	26%
FBD [21]	0.70 (0.21)	77 (20)	92 (11)	93	40%
CNNpre [21]	0.68 (0.22)	80 (22)	90 (12)	92	28%
Proposed CNN	**0.76 (0.22)**	**90 (22)**	**88 (16)**	**95**	-

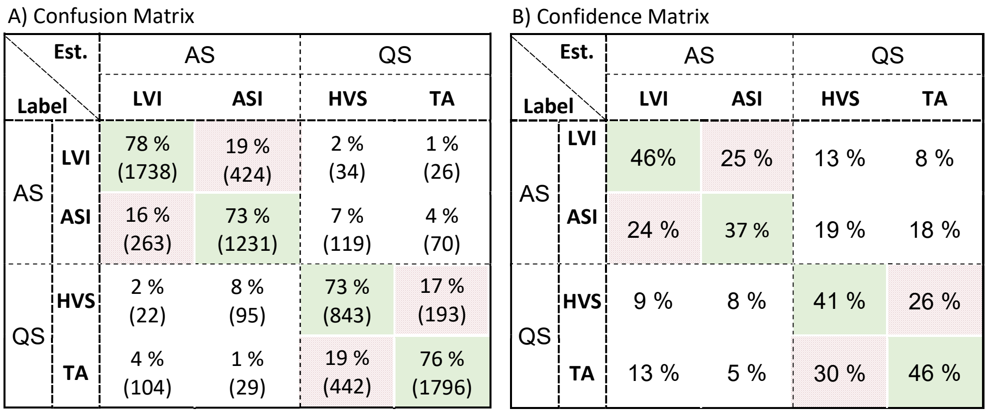
Sens: sensitivity, Spec: specificity, AUC: area under the mean ROC curves, p_{boot}: proportion of the bootstrapped recordings (in percent) when the model is resulting in a better performance than the proposed method.

Table II
The classification performance of the proposed methods compared to the state-of-the-art for the QS detection in the term recordings

Model	Mean Kappa (SD)	Mean % Sens (SD)	Mean % Spec (SD)	p_{boot}
CLASS [19]	0.62 (0.19)	95 (06)	86 (06)	0%
HMM [20]	0.82 (0.19)	89 (19)	94 (09)	32%
HMM$_R$ [20]	0.89 (0.07)	94 (04)	96 (04)	38%
GMM [20]	0.81 (0.19)	89 (19)	93 (06)	38%
GMM$_R$ [20]	0.85 (0.12)	92 (06)	93 (06)	25%
Proposed CNN	**0.91 (0.07)**	**95 (05)**	**96 (04)**	-

Table III
The classification performance of the proposed methods compared to the state-of-the-art for the term 4-class classification

Model	Mean Kappa (SD)	Mean % Sens (SD)	Mean % Spec (SD)	p_{boot}
HMM [20]	0.54 (0.20)	62 (13)	90 (05)	19%
HMM$_R$ [20]	0.61 (0.10)	71 (06)	91 (02)	19%
GMM [20]	0.51 (0.16)	60 (11)	87 (07)	19%
GMM$_R$ [20]	0.51 (0.11)	63 (09)	88 (06)	6%
Proposed CNN	**0.66 (0.14)**	**72 (12)**	**92 (03)**	-



A) Confusion Matrix

Est.		AS		QS	
Label		LVI	ASI	HVS	TA
AS	LVI	78 % (1738)	19 % (424)	2 % (34)	1 % (26)
	ASI	16 % (263)	73 % (1231)	7 % (119)	4 % (70)
QS	HVS	2 % (22)	8 % (95)	73 % (843)	17 % (193)
	TA	4 % (104)	1 % (29)	19 % (442)	76 % (1796)

B) Confidence Matrix

Est.		AS		QS	
Label		LVI	ASI	HVS	TA
AS	LVI	46%	25 %	13 %	8 %
	ASI	24 %	37 %	19 %	18 %
QS	HVS	9 %	8 %	41 %	26 %
	TA	13 %	5 %	30 %	46 %

Figure 3: Confusion matrix (A) and Confidence matrix (B) for the proposed method for 4 sleep stage classification in the term dataset

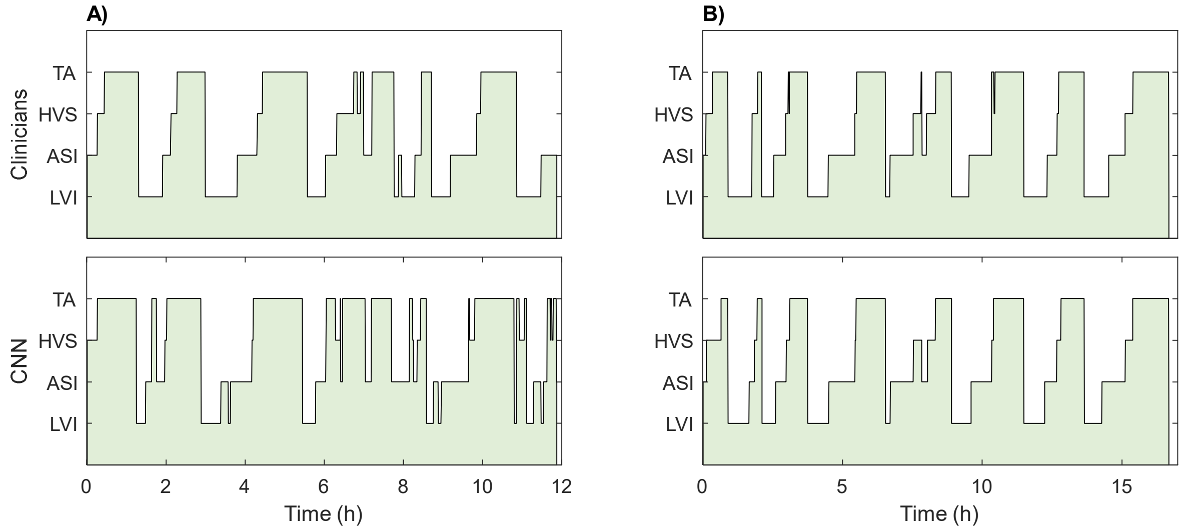


Figure 4: Hypnogram of two term recordings for which the CNN performance is A) poor (kappa = 0.36) and B) almost perfect (kappa = 0.87). The top row corresponds to the clinicians' labels and the bottom corresponds to the proposed CNN outputs.

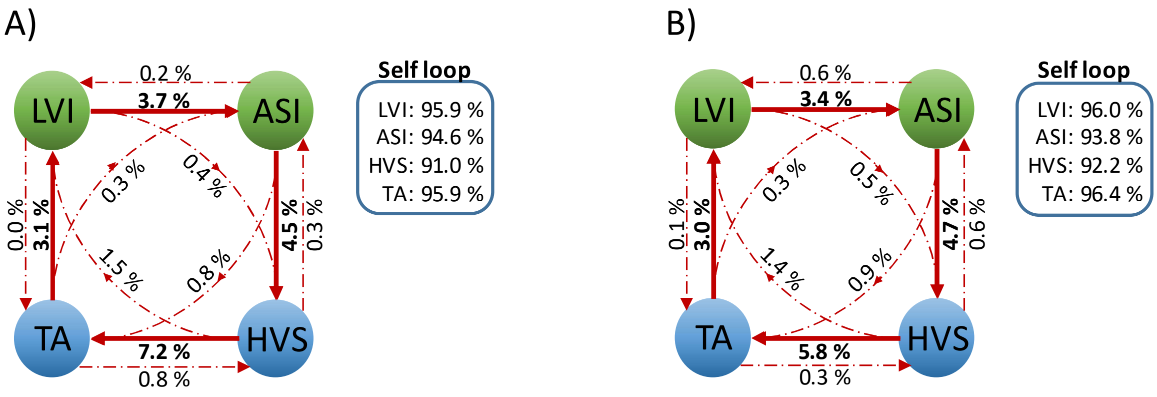


Figure 5: Transition matrix for different sleep states based on the clinicians' labels (A) and the smoothed CNN outputs (B). The percentages shown on the right side of each graph define the probability of remaining in the same state.

The transition graphs for the clinicians' labels are shown in Figure 5 (A) and the proposed model in Figure 5 (B). These are averaged across all the recordings. Each node of these graphs has 3 outward links and 3 inward links representing the probabilities of transitions between the states. On the right side of each graph, the self-loop edges defining the probability of remaining in the same state are listed. In each node, the summation of the self-loop and outward probabilities should

Table IV
Mean Kappa for the proposed methods when it is fed and retrained by different number of EEG channels. The numbers in the parentheses represent the standard deviation.

		Preterm 2-class classification		Term 4-class classification	
N[*]	Used channels	kappa	p_{boot}	kappa	p_{boot}
1	C3 - C4[**]	0.60 (0.31)	16%	0.44 (0.16)	6%
2	C3, C4	0.62 (0.25)	28%	0.41 (0.16)	6%
4	C3, C4, O1, O2	0.67 (0.24)	19%	0.50 (0.19)	12%
8	All	0.76 (0.22)	-	0.66 (0.14)	-

[*] Number of used channels
[**] Subtraction of C4 from C3 (bipolar) as this is the input of the CFM

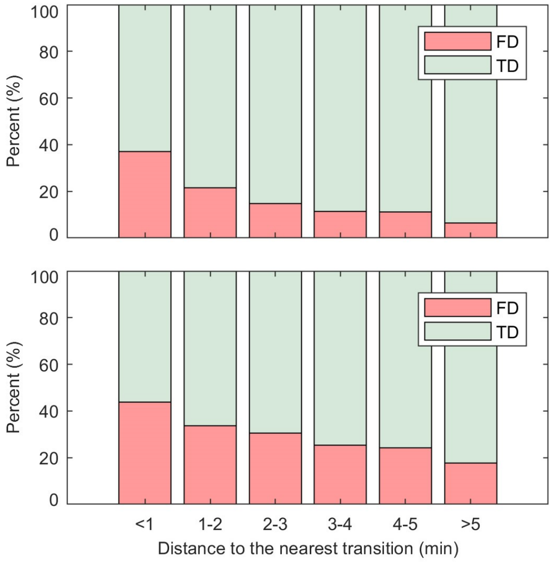


Figure 6: False and true detection rates as a function of the distance to the nearest sleep state transition. The top and bottom rows respectively correspond to the test results from the 2-class and 4-class classification problem. In this figure, the red and green colours represent the falsely detected (FD) and truly detected (TD) segments, respectively.

equal 100% (± 0.1% due to the rounding effect).Thicker links represent the most likely transitions.

Table IV lists the mean Kappa and the SD when the input of the CNN has a different number of EEG channels (dropping and retraining) for both 2-class and 4-class classification. In each case, the CNN has been retrained multiple times with different random initializations and the model performing the best on the validation data has been taken to maximize the efficiency. Note that QS detection in term neonates is not considered in this channel analysis, since it uses the same CNN that was trained for preterm QS detection.

In the second analysis using a fixed CNN without retraining, when only one channel is dropped, the kappa range is 0.47-0.76 (median: 0.71) in preterm and 0.41-0.67 (median: 0.61) in term datasets depending on the dropped channels. These values drop to 0.11-0.74 (median: 0.69) and 0.11-0.64 (median: 0.52) respectively if 2 channels are detached. In case of removing more than 2 channels, the CNN cannot perform reliably in either preterm or term classification.

Furthermore, the overlap ratio of the input segments and the size of the averaging filter in the post-processing have been analysed. Increasing the overlap ratio from 0% to 25%, 50% and 75% has no meaningful effect on the CNN training. However, raising this ratio increases the number of segments, and consequently the number of estimated probabilities, for each recording, which leads to a smoother output. For instance, 3 minutes of EEG is split into 6 segments with no overlap (totalling 6 probabilities for 3 minutes), while it can be split into 12 segments with 50% overlap (totalling 12 probabilities for 3 minutes). This results in a slight increase of kappa (by 0.01) in the post-processing smoothing step. The length of the smoothing filter has also considered between 0 and 30 points. No significant difference in averaged performance is observed when this length varies between 4 and 10. Therefore, the length of 6 has been chosen as was taken in the previous study [21] based on physiological reasons.

C. *False detections*

In Figure 6, the ratios of wrongly and correctly classified segments are shown with respect to the distance between the segment and the nearest sleep state transition. In this figure, each group is expressed in percentage to be independent from the number of samples. For instance, in both 2-class (top row) and 4-class (bottom row) classification problems, almost 40% of the segments that are in the range of ±1 minute from the sleep transitions were classified wrongly. However, this ratio changes to 6% (2-class) and 18% (4-class) for the segments that are far from the transitions (>5 min).

In total, 24% (2-class) and 17% (4-class) of all false detections occurred in one minute before or after the transitions.

IV. Discussion

In this study, a new design of CNN with more complexity and generalization power has been proposed for 2-class classification in preterms and in term babies, as well as 4-class classification in term infants. The results have been compared with state-of-the-art methods that have been trained and tested on the same database. As was shown in Tables I to III, the proposed CNN outperforms the compared methods in all three problems. One important advantage of the proposed method is that it can be used in semi real-time applications. It is 'semi' because of the smoothing filter in the post-processing step, which causes a delay of 3 minutes. However, the considered HMM and GMM with rescaling, which have a reduced performance compared to the proposed method, requires the whole EEG recording to estimate the class labels, and, therefore, they cannot be used in real-time applications. The high performance of the proposed method, in addition to its semi real-time characteristics, makes this method a good candidate for real-time NICU brain monitoring.

The confusion matrix shows that the main challenge in neonatal sleep staging is the detection of sub-states. Although QS and AS can be classified accurately, discrimination between QS-TA and QS-HVS, as well as between AS-ASI and AS-LVI, is more challenging. The confidence matrix revealed that the falsely detected segments, including false intra-state detections (TA vs. HVS and ASI vs. LVI), have meaningfully lower confidence compared to the correctly detected segments. It shows that when a segment is classified correctly, the network makes a big margin between the probabilities of the true class and the wrong ones. This leads to a high robustness. On the other hand, there is a significantly smaller confidence margin for the false detections. This confidence analysis can be useful for further improvements in the future.

The false detection ratios, in Figure 6, demonstrate that another challenge, in both 2-class and 4-class problems, is the classification of the epochs that are close to the sleep transitions. There are two possible explanations for this performance reduction near the sleep transition. The first is in the nature of the sleep staging in which the transitionary sleep states have indistinct patterns even for human experts [8]. The second is the effect of the moving-average filter used in the post-processing step. Although this filter improves the total performance, it smooths all transitions and decreases the accuracy of sharp transition detections. Developing an advanced technique to predict sleep transitions to adaptively adjust the smoothness can reduce this false detection rate. However, due to the presence of uncertain transitional sleep state, it does not seem to have important clinical added value. Furthermore, analysis on the recordings showed that in the used preterm dataset, there are two recordings having poor EEG quality or intravenous infusion motor artefact that resulted in many false detections, which seems unlikely to be classifiable without the domain knowledge in such a network.

In the term neonatal sleep staging, the expected transition cycle is ASI → HVS → TA → LVI → ASI (omitting the transitional sleep states) [8], as shown in the transition graph of Figure 5. This graph also displayed that the CNN accurately follows this transition cycle. Although a big effort was made to advance the proposed method with a long short-term memory (LSTM) in order to learn such a transition by a data–driven approach (instead of the current smoothing filter), as is shown useful for adults [35], no further meaningful improvement was achieved.

While not every NICU has the capacity to record continuous EEG, with 8 channels (or more), we also investigated the performance when only a few channels are recorded. These channels are selected based on the inputs of CFMs with 2 or 4 channels which are commonly used in some medical centres. In this analysis, it was shown that in both preterm 2-class and term 4-class classification, increasing the number of available EEG channels leads to a higher performance. Furthermore, as mentioned, the performance of the network drops if more than 2 channels are detached during the testing phase. Although in some particular channel configurations the detachment has no big effect on the performance, an average random detachment results in a high reduction in performance.

In this study, there are two limitations that should be taken into consideration: first, only the algorithms that were available for us have been compared with the proposed method. The reason behind it is the fact that comparing various algorithms that are trained and validated on different datasets with different characteristics (e.g. neonatal population with dissimilar PMA or GA, labellers, number of electrodes, presence of artefacts, and standards of EEG recording) leads to a biased and incorrect conclusion. Second, in this study, two annotators from the same centre labelled the sleep stages. Hence, one might consider it imperfect compared to multi-centre studies with multiple annotators that should be done in future.

As future work, we aim to use multimodal inputs to further improve the classification tasks. Furthermore, extra considerations should be taken into account during the network design and training procedure in order to attain a better performance with fewer channels of EEG, which can be important for practical applications. In addition, due to not having an appropriately big term dataset, the structure of the network was not optimized for these babies. With more labelled data at this age, the structural optimization would be expected to improve the performance further. Besides, a multi-centre multi-rater study can validate the performance of the proposed method under different practical conditions.

V. Conclusion

In this study, we present a new architecture of CNN that can detect sleep states in both preterm and term neonates. This end-to-end model uses raw multichannel EEG data as input and results in class probabilities, without the need for any complicated pre/post-processing or any prior hand-engineered feature extraction. For 2-class classification, the high performance of this method and its ability to operate in almost real-time makes it the first option to be used in the neonatal brain monitor developed in our centre, namely NeoGuard.

Acknowledgements

The Tesla K40 GPU used for this research was donated by the NVIDIA Corporation.

References

[1] P. Garg, R. Krishak, and D. K. Shukla, “NICU in a community level hospital,” *Indian J. Pediatr.*, vol. 72, no. 1, pp. 27–30, Jan. 2005.

[2] S. Fallah, X.-K. Chen, D. Lefebvre, J. Kurji, J. Hader, and K. Leeb, “Babies admitted to NICU/ICU: Province of birth and mode of delivery matter,” *Heal. Q*, vol. 14, no. 2, pp. 16–20, 2011.

[3] K. A. Ziegler, D. A. Paul, M. Hoffman, and R. Locke, “Variation in NICU Admission Rates Without Identifiable Cause,” *Hosp. Pediatr.*, vol. 6, no. 5, pp. 255–260, May 2016.

[4] H.-C. Kung, D. L. Hoyert, J. Xu, and S. L. Murphy, “Deaths: final data for 2005,” *Natl Vital Stat Rep*, vol. 56, no. 10, pp. 1–120, 2008.

[5] H. Blencowe *et al.*, “Born Too Soon: The global epidemiology of 15 million preterm births,” *Reprod. Health*, vol. 10, no. 1, p. S2, Nov. 2013.

[6] M. J. Benders *et al.*, “Early Brain Activity Relates to Subsequent Brain Growth in Premature Infants,” *Cereb. Cortex*, vol. 25, no. 9, pp. 3014–3024, Sep. 2015.

[7] K. K. Iyer *et al.*, “Cortical burst dynamics predict clinical outcome early in extremely preterm infants,” *Brain*, vol. 138, no. 8, pp. 2206–2218, Aug. 2015.

[8] A. Dereymaeker *et al.*, “Review of sleep-EEG in preterm and term neonates,” *Early Hum. Dev.*, vol. 113, pp. 87–103, 2017.

[9] 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.*, vol. 12, no. 1, pp. 58–70, 2009.

[10] K. A. Allen, “Promoting and Protecting Infant Sleep,” *Adv. Neonatal Care Off. J. Natl. Assoc. Neonatal Nurses*, vol. 12, no. 5, pp. 288–291, Oct. 2012.

[11] D. Y. Barbeau and M. D. Weiss, "Sleep Disturbances in Newborns," *Children*, vol. 4, no. 10, p. 90, Oct. 2017.

[12] N. Koolen *et al.*, "Automated classification of neonatal sleep states using EEG," *Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol.*, vol. 128, no. 6, pp. 1100–1108, 2017.

[13] A. Piryatinska, G. Terdik, W. A. Woyczynski, K. A. Loparo, M. S. Scher, and A. Zlotnik, "Automated detection of neonate EEG sleep stages," *Comput. Methods Programs Biomed.*, vol. 95, no. 1, pp. 31–46, Jul. 2009.

[14] K. Palmu, T. Kirjavainen, S. Stjerna, T. Salokivi, and S. Vanhatalo, "Sleep wake cycling in early preterm infants: Comparison of polysomnographic recordings with a novel EEG-based index," *Clin. Neurophysiol.*, vol. 124, no. 9, pp. 1807–1814, Sep. 2013.

[15] N. J. Stevenson, K. Palmu, S. Wikström, L. Hellström-Westas, and S. Vanhatalo, "Measuring brain activity cycling (BAC) in long term EEG monitoring of preterm babies," *Physiol. Meas.*, vol. 35, no. 7, pp. 1493–1508, Jun. 2014.

[16] J. S. Barlow, "Computer characterization of tracé alternant and REM sleep patterns in the neonatal EEG by adaptive segmentation—an exploratory study," *Electroencephalogr. Clin. Neurophysiol.*, vol. 60, no. 2, pp. 163–173, Feb. 1985.

[17] 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*, 2005, pp. 6014–6016.

[18] V. Krajča, S. Petránek, J. Mohylová, K. Paul, V. Gerla, and L. Lhotská, "Neonatal EEG Sleep Stages Modelling by Temporal Profiles," in *Computer Aided Systems Theory – EUROCAST 2007*, 2007, pp. 195–201.

[19] A. Dereymaeker *et al.*, "An Automated Quiet Sleep Detection Approach in Preterm Infants as a Gateway to Assess Brain Maturation," *Int. J. Neural Syst.*, vol. 27, no. 6, p. 1750023, Sep. 2017.

[20] K. Pillay, A. Dereymaeker, K. Jansen, G. Naulaers, S. Van Huffel, and M. De Vos, "Automated EEG sleep staging in the term-age baby using a generative modelling approach," *J. Neural Eng.*, vol. 15, no. 3, p. 036004, Jun. 2018.

[21] A. H. Ansari *et al.*, "Quiet sleep detection in preterm infants using deep convolutional neural networks," *J. Neural Eng.*, vol. 15, no. 6, p. 066006, 2018.

[22] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *ArXiv150203167 Cs*, Feb. 2015.

[23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[24] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, vol. 1. MIT press Cambridge, 2016.

[26] A. H. Ansari, P. Cherian, A. Caicedo Dorado, G. Naulaers, M. De Vos, and S. Van Huffel, "Neonatal Seizure Detection Using Deep Convolutional Neural Networks," *Int. J. Neural Syst.*, no. accepted, 2018.

[27] A. O'Shea, G. Lightbody, G. Boylan, and A. Temko, "Neonatal Seizure Detection using Convolutional Neural Networks," *ArXiv Prepr. ArXiv170905849*, 2017.

[28] 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.*, vol. 100, pp. 270–278, 2018.

[29] H. Cecotti and A. Graeser, "Convolutional neural network with embedded fourier transform for EEG classification," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, 2008, pp. 1–4.

[30] H. Cecotti and A. Graser, "Convolutional neural networks for P300 detection with application to brain-computer interfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 433–445, 2011.

[31] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. De Vos, "Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification," *IEEE Trans. Biomed. Eng.*, pp. 1–12, Accepted 2018.

[32] A. Vilamala, K. H. Madsen, and L. K. Hansen, "Deep convolutional neural networks for interpretable analysis of EEG sleep stage scoring," in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2017, pp. 1–6.

[33] A. Sors, S. Bonnet, S. Mirek, L. Vercueil, and J.-F. Payen, "A convolutional neural network for sleep stage scoring from raw single-channel EEG," *Biomed. Signal Process. Control*, vol. 42, pp. 107–114, Apr. 2018.

[34] J. Zhang and Y. Wu, "Automatic sleep stage classification of single-channel EEG by using complex-valued convolutional neural network," *Biomed. Eng. Biomed. Tech.*, vol. 63, no. 2, pp. 177–190, 2018.

[35] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 3, pp. 400–410, Mar. 2019.

[36] R. T. Schirrmeister *et al.*, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapp.*, vol. 38, no. 11, pp. 5391–5420, 2017.

[37] L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "EEG-based detection of the attended speaker and the locus of auditory attention with convolutional neural networks," *bioRxiv*, p. 475673, Dec. 2018.