Sleep State Trend (SST), a bedside measure of neonatal sleep state fluctuations based on single EEG channels

Montazeri Moghadam, Saeed

2022-11

Sleep State Trend (SST), a bedside measure of neonatal sleep state fluctuations based on single EEG channels

Saeed Montazeri Moghadam [a,b,*], Päivi Nevalainen [a], Nathan J. Stevenson [c], Sampsa Vanhatalo [a,b]

[a] BABA Center, Department of Clinical Neurophysiology, HUS diagnostic center, Children's Hospital, University of Helsinki and Helsinki University Hospital, Helsinki, Finland
[b] Department of Physiology, University of Helsinki, Helsinki, Finland
[c] Brain Modeling Group, QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia

HIGHLIGHTS

- Monitoring sleep wake cycling is an essential component in neonatal brain monitoring.
- Detection of quiet sleep epochs is achievable from single EEG channels with deep learning-based methods.
- Sleep State Trend (SST) can be used to visualize the classifier results in bedside monitors.

ABSTRACT

Objective: To develop and validate an automated method for bedside monitoring of sleep state fluctuations in neonatal intensive care units.
Methods: A deep learning-based algorithm was designed and trained using 53 EEG recordings from a long-term (a)EEG monitoring in 30 near-term neonates. The results were validated using an independent dataset from 30 polysomnography recordings. In addition, we constructed Sleep State Trend (SST), a bedside-ready means for visualizing classifier outputs.
Results: The accuracy of quiet sleep detection in the training data was 90%, and the accuracy was comparable (85–86 %) in all bipolar derivations available from the 4-electrode recordings. The algorithm generalized well to a polysomnography dataset, showing 81% overall accuracy despite different signal derivations. SST allowed an intuitive, clear visualization of the classifier output.
Conclusions: Fluctuations in sleep states can be detected at high fidelity from a single EEG channel, and the results can be visualized as a transparent and intuitive trend in the bedside monitors.
Significance: The Sleep State Trend (SST) may provide caregivers and clinical studies a real-time view of sleep state fluctuations and its cyclicity.

© 2022 International Federation of Clinical Neurophysiology. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Sleep is essential for early brain organization (Allen, 2012, Graven, 2006, Marks et al., 1995, Roffwarg et al., 1966), and a large number of studies have shown the impact of sleep on many levels of neurobehavioral development (Arditi-Babchuk et al., 2009, Ednick et al., 2009, Lam et al., 2003, Mirmiran et al., 2003, Scher et al., 1996, Touchette et al., 2007, van den Hoogen et al., 2017). Clinically, fluctuation of sleep and wake states is seen as a key, holistic indicator of healthy brain function, hence monitoring sleep-wake rhythms has become a common practice in early brain monitoring in the neonatal intensive care units (NICU) (Meder et al., 2022, Thoresen et al., 2010, van den Hoogen et al., 2017). It has also been suggested that sleep states should be monitored at the bedside to help in optimizing care procedures in the interest of minimizing disruption during a particular sleep state, or the sleep-wake cycling (SWC) (Dereymaeker et al., 2017a).

The gold standard of sleep state recognition is behavioral observation (Grigg-Damberger et al., 2007), however physiological recordings can also be used and polysomnography (PSG) recording is commonly taken as the gold standard physiological recording method. It includes electroencephalography (EEG) and polygraphy (respiration, muscle tone, eye movements) signals, implying technical challenges, especially in bedside NICU use. Therefore, methods have been developed to allow an automated sleep

* Corresponding author at: Department of Physiology, Biomedicum 1, Room B129b, University of Helsinki, Haartmaninkatu 8, 00290 Helsinki, Finland.
E-mail address: saeed.montazeri@helsinki.fi (S. Montazeri Moghadam).

S. Montazeri Moghadam, Päivi Nevalainen, N.J. Stevenson et al.

Clinical Neurophysiology 143 (2022) 75–83

assessment from the EEG signal alone (Ansari et al., 2018, Ansari et al., 2020, Ansari et al., 2022, Dereymaeker et al., 2017b, Fraiwan and Alkhodari, 2020, Fraiwan et al., 2011, Ghimatgar et al., 2020, Hsu et al., 2013, Koolen et al., 2017, Pillay et al., 2018, Piryatinska et al., 2009). These methods usually aim at recognizing the two cardinal sleep states of a neonate, active sleep (AS or rapid eye movement sleep, REM) and quiet sleep (QS or non-REM sleep); some works have even aimed at recognizing more sleep states, including separate detection of the wake state (Ansari et al., 2020; Fraiwan et al., 2011; Fraiwan and Alkhodari, 2020; Hsu et al., 2013; Pillay et al., 2018). Most of the existing sleep state detectors are trained using feature-based classifiers (Dereymaeker et al., 2017b; Fraiwan et al., 2011; Ghimatgar et al., 2020; Hsu et al., 2013; Koolen et al., 2017; Piryatinska et al., 2009), and they generally show performance that comes fairly close to PSG detection. However, reasonably high classifier performance has required use of multiple EEG signals and/or temporal smoothing, both of which challenge the utility of such solutions in the typical limited-channel EEG monitoring that prevails in the NICUs worldwide. Some studies using convolutional neural networks (CNNs) have shown improvement in automatic neonatal sleep state classification, even with fewer EEG channels (Ansari et al., 2018, Ansari et al., 2020, Ansari et al., 2022).

Here, we aimed to develop a sleep state classifier that meets the essential needs of NICU implementation, i.e. the classifier should recognize EEG between active and quiet sleep with high enough accuracy using a single channel EEG only; further, we purposed to construct a classifier that would allow an intuitively interpretable visualization, a Sleep State Trend (SST), to be implemented in the bedside EEG monitors. To this end, we employed convolutional neural networks (CNNs) capable of working on any number of EEG channels, and we trained and tested it using both EEG and PSG recordings. In addition, we tested the performance of our novel classifier against a reference classifier described in (Koolen et al., 2017).

2. Materials and methods

2.1. Overview

An overview of the development and performance testing of the sleep state classifier and the SST is shown in Fig. 1.

2.2. EEG datasets

Cross-validation dataset: We used a cohort of **30** neonates with a total of **943.7** hours of recordings (13 females, gestational age (GA) between 35 + 1 and 42 + 1 weeks), all of which exhibited clear sleep-wake cycling in the amplitude-integrated electroencephalography (aEEG) trend. These recordings were collected retrospectively from a clinical database of long-term monitoring (Helsinki University Children's Hospital, Finland) (Nevalainen et al., 2021, Nevalainen et al., 2019). EEGs were collected using a NicoletOne system (Cardinal Healthcare/Natus, USA) at 256 Hz with four scalp electrodes located at F3, F4, P3, and P4, based on the international 10–20 system (Cherian et al., 2009).

The continuous EEG signal was pre-processed with an automated pipeline. First, all the bipolar EEG channels (F3-P3, F4-P4, F3-F4, and P3-P4) were scanned automatically to find and discard high-amplitude values (> ±250 μV) or flat signals (constant zero). In total 118.6 hours of the signal were detected as artifacts and rejected from further analysis. Third, each EEG signal was band-pass filtered at 0.5–30 Hz with a 4th order Butterworth filter. Fourth, signals were resampled to 64 Hz with an anti-aliasing filter

and segmented into 1-minute non-overlapping epochs (discrete signal length of 3840 samples).

This dataset was annotated visually by two experts (SV and PN; hereafter referred to as E1 and E2) for only QS epochs. Since there was only EEG available, the score QS was assigned to epochs with clearly discontinuous, or tracé alternant background pattern. Conversely, the rest was taken as continuous EEG, which may be either AS or awake. This dichotomic scoring may not match perfectly with the vigilance state scoring obtained from a full PSG study, however an approach of this kind has been used in many recent studies, and it is also the underlying assumption in the widely used SWC assessment of the aEEG trends (Thoresen et al., 2010). Notably, the annotations were done at one second accuracy for all the EEG considered to represent QS, i.e. no fixed segment lengths were used. This approach allows physiologically faithful annotation of the QS periods, although it deviates from the clinical tradition of using fixed epoch lengths in the sleep scoring. In total, 175.6 hours of QS (median per neonate: 21 minutes, interquartile range (IQR): 15–27 minutes) were annotated by E1 and 243 hours of QS (median per neonate: 25 minutes, IQR: 18–33 minutes) were annotated by E2. Every 1-minute length EEG epoch was labelled into one of QS or AS scores if all the samples in the epoch belong to the same QS or AS score.

Independent validation dataset: The algorithm was validated using an independent dataset that included **30** infants (12 female, GA: 30 + 3 – 41 + 1 weeks) with PSG recordings, the gold standard in sleep state classification. These recordings were done using Embla N700 equipment and RemLogic 3.2.0 software (Natus, United States) as per routine clinical protocol, and the dataset was collected retrospectively from the hospital archives (Helsinki University Children's Hospital, Finland) using following criteria: the PSG study was performed during newborn period (conceptional age less than 45 weeks), clinical interpretation was normal, and the EEG exhibited tracé alternant as the dominant form of EEG during quiet sleep. Such dataset includes four channels of EEG (C4, O2, A1, Fz) which are not the same recording positions as the training dataset (see above), and hence allow testing how well the findings generalize to EEG in different scalp locations. In addition, the PSG data includes electromyograms (EMGs), electrooculograms (EOGs), electrocardiogram (ECG) and respiratory signals.

The continuous EEG signal was pre-processed with an automated pipeline. First, all the bipolar EEG channels (C4-O2, O2-A1, C4-A1, C4-Fz) were scanned automatically to find and discard high-amplitude values (> ±250 μV) or flat signals (constant zero). Then, each EEG signal was band-pass filtered at 0.5–30 Hz with a 4th order Butterworth filter. Finally, signals were resampled to 64 Hz with an anti-aliasing filter and segmented into 1-minute non-overlapping epochs.

The PSG dataset was annotated by clinical sleep specialists as a part of their clinical work, employing the sleep state scoring as per international guidelines established by the American Academy of Sleep Medicine (AASM) (Grigg-Damberger et al., 2007). For the present work, we exported their hypnogram annotations including all the sleep states recognized in our neonatal sleep studies (wake, REM, N1, N3). Here, we tested the algorithm first with a two-class classification where QS was taken to correspond to N3 in the PSG hypnograms, and AS was taken from the other sleep states (wake, REM, N1) grouped together. Notably, such PSG studies include a minimal amount of actual wake, hence the grouping was primarily for REM and N1 states. We also assessed an alternative classification where wake, REM, N1 and N3 were considered separately.

Institutional Research Review Board of the HUS diagnostic center approved the study, including waiver of consent due to the retrospective and observational nature of the study.
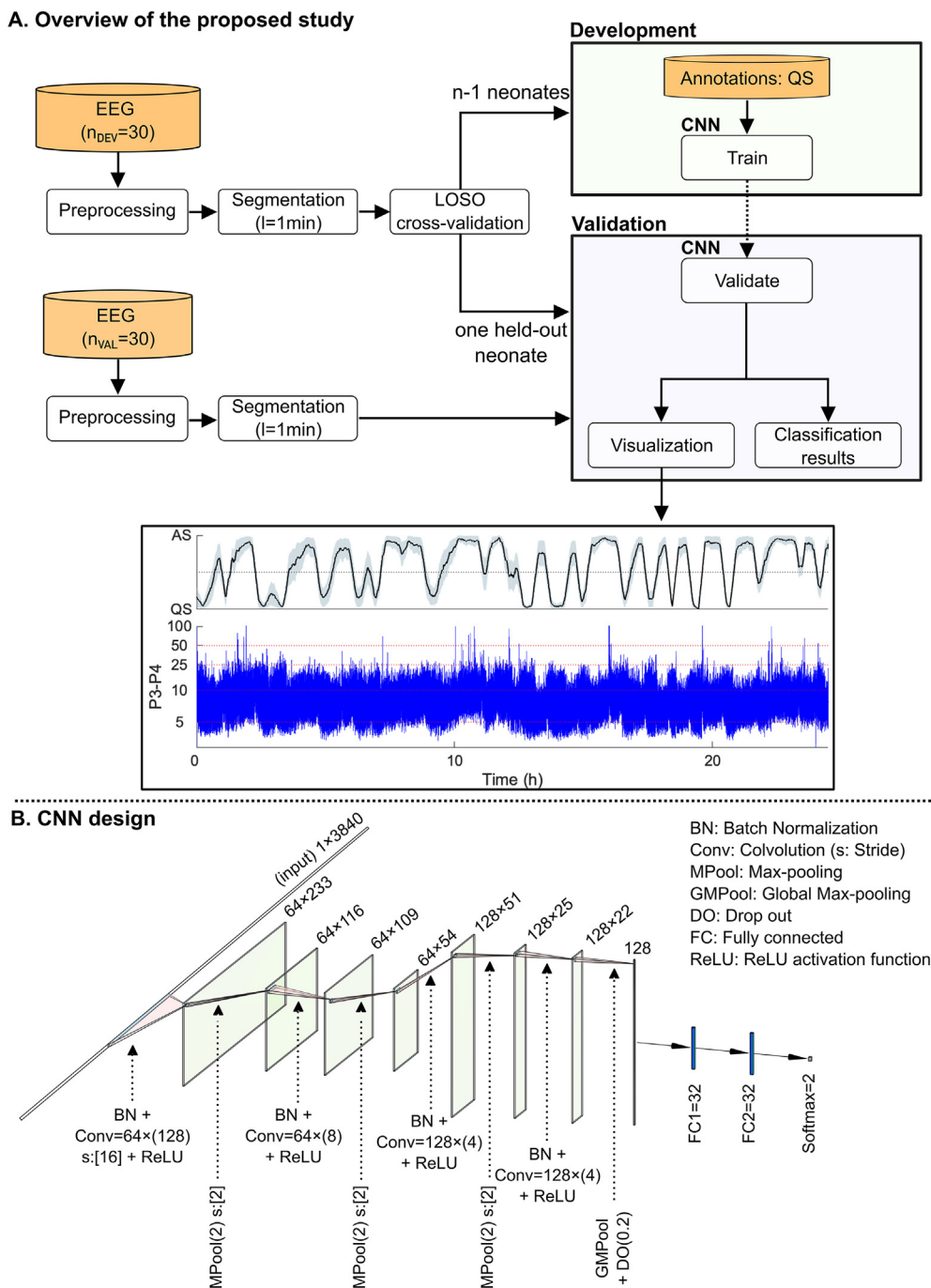
A. Overview of the proposed study

B. CNN design

Fig. 1. A. The overview of the sleep state classifier training, its performance assessment, validation, and visualization. **B**. Layer description of the proposed CNN for single-channel sleep state classification with a total of 5,082 parameters. LOSO: Leave-one-subject-out, QS: Quiet sleep, CNN: Convolutional Neural Network.

2.3. Algorithm design

A CNN is a type of artificial neural network (ANN) that is composed of interchanging layers of convolution, nonlinear operator, and pooling. A convolution layer convolves the n-dimensional input tensor and weights matrix. A nonlinear operator (most commonly ReLU) is then used after each convolution layer to make a nonlinear hidden layer for the CNN. A pooling layer performs a down-sampling process on the output volume of its previous layer to control overfitting. These layers automatically extract features. Then, usually fully connected layers follow those layers to perform the classification. An additional type of layers may be applied

depending on the problem, such as batch-normalization to standardize the input to a layer (Ioffe and Szegedy, 2015), drop-out to help prevent overfitting and increase generalization (Srivastava et al., 2014), and softmax to map the output of the last layer to a probability distribution (Bishop, 1995), etc.

The proposed CNN design in this study receives a single-channel 1-minute EEG segment which is initially resampled to 64 Hz as input and after 11 layers of processing gives a vector of probabilities for AS and QS classes as output. Fig. 1B illustrates the proposed CNN structure which has 5,082 parameters. Prior work has shown that CNN needs a sufficient number of parameters for accurate generalization (Bubeck and Sellke, 2021). The architec-

S. Montazeri Moghadam, Päivi Nevalainen, N.J. Stevenson et al.

Clinical Neurophysiology 143 (2022) 75–83

ture of the network was designed based on some previous studies (Ansari et al., 2018, Ansari et al., 2020, Ansari et al., 2022, Phan et al., 2019a, 2019b, Sors et al., 2018, Zhang and Wu, 2018, Zhang et al., 2020), and trial and error using the development data with a focus on accepting a single-channel EEG as input. Compared to the 1D-CNN solution in prior study, our design has higher number of parameters (Ansari et al., 2018). The implementation was done in Python using Keras with a TensorFlow backend and trained on a Geforce GTX 1070 GPU. The first 8 layers perform feature extraction, and the last 3 layers perform classification and probability calculation.

The optimal CNN classifier was trained with an ADAM solver (beta1 = 0.9, beta2 = 0.99, and learning rate = 0.001) and minibatch size of 64. The parameters were randomly initialized using uniform He initialization (He et al., 2015). Training continued until the validation loss stopped decreasing for a period of at least 35 epochs. Maximum training epochs were set at 500, while most trainings stopped before 100. The learning rate is reduced by the factor of 0.1 once learning plateaus over 20 consecutive epochs.

Post-processing was used for smoothing out spatial and temporal noise in the classifications. The CNN generates an output class probability for each processed 1-minute EEG segment from each channel. This temporal resolution is far higher than what is needed for the NICU brain monitoring purposes where sleep state cycling occurs in the scale of tens of minutes (Curzi-Dascalova et al., 1988, Osredkar et al., 2005, Stevenson et al., 2014) For spatial smoothing, we combined the output probabilities across channels with an averaging function followed by ArgMax to determine the sleep state for each 1 min epoch. A very mild temporal smoothing (moving median filter, window length 5 epochs, i.e. 5 minutes) was applied to the combined CNN output in order to taper down incidental noise in the time series.

In the following, we will describe channel-wise results from all four EEG channels separately after the mild temporal smoothing. There was no major difference between the examined four EEG derivations, and each derivation alone gives reasonably accurate results. Channel-wise results are shown to provide a solution for all EEG monitoring contexts, however our post-hoc experiment indicated that spatial smoothing would improve classifier performance (Supplementary Fig. S1 and Supplementary Table S1).

2.4. Training and performance testing

Classifier training was done with annotations from both experts; thus, information was incorporated from both the agreements and disagreements, which represent complementary aspects of the experts' annotations, and may improve classifier performance (Moghadam et al., 2021). We excluded epochs with multiple sleep states. The classifier produces one list of two estimated class probabilities (output of the Softmax layer) for every 1-minute non-overlapping EEG epoch from a single channel. For each epoch, the class with higher probability was taken as the prediction.

Classifier performance was initially tested with the training dataset using leave-one-subject-out (LOSO) cross-validation approach, i.e. training with *n-1* neonates and testing on the one "held-out" neonate. This was repeated for all *n* folds. As part of the training process, to prevent overfitting, 10 % of the data in the training folds of LOSO was set aside as inner validation (training was stopped if the inner validation loss did not decrease after at least 20 epochs).

For the metrics of classifier performance, we computed confusion matrices (true positive (TP) = correct detection of QS epochs), accuracy, precision, F1 score, Cohen's kappa, and the receiver operator characteristics (ROC, in Supplementary Material) curve. These metrics can be defined with the numbers of true positives/ QS (TP), true negatives/AS (TN), false positives (FP) and false negatives (FN).

Accuracy is defined by (TP + TN)/(TP + TN + FP + FN), precision by TP/(TP + FP), F1 score by $(2 \times TP)/(2 \times TP + FP + FN)$ and Cohen's kappa coefficient by $k = (P_o - P_e)/(1 - P_e)$, where P_o represents the observed agreement and P_e denotes the chance agreement.

Clinical utility in individual diagnostics was estimated from the ranges (min–max) of classifier performance measures in individual neonates.

2.5. Visualization of the Sleep State Trend (SST)

Bedside implementation of the developed system requires an intuitive and transparent visualization of the classifier output. As a first step toward such visualization, the clinician should be able to see the classifier output alongside an estimate of its certainty, to be informed of, e.g., ambiguity in the EEG signal for biological or technical reasons. To this end, we propose Sleep State Trend (SST) by taking a weighted average of the 'probability' outputs from the sleep state classifier for every 1-min EEG epoch. In addition to the SST, an index of classifier uncertainty is added to depict the distribution of classifiers outputs from all channels available in the given recording.

The SST algorithm and its visualization can be used openly via a cloud service by requesting access from the authors.

2.6. Comparison to a reference classifier

A recently published feature-based classifier (Koolen et al., 2017) was used as a reference classifier to provide a benchmark with the proposed novel classifier. This allowed comparison of a feature-based approach with our proposed end-to-end deep-learning approach. We also compared our results with this reference classifier after re-training it on our present dataset to the classifier trained earlier with preterm EEG data (67 infants, (Koolen et al., 2017)). Thus, we could assess how well the sleep state detection algorithms could generalize between preterm and term infants. The Koolen's classifier utilizes a set of N = 57 computational features combined in a support vector machine (SVM) classifier. The features are extracted from 10-min EEG segments of multichannel EEG, including following categories: (1–2) age-related features, (3–12) frequency-domain features, (13–47) time-domain features, and (48–57) spatial connectivity features. Thus, each multichannel EEG segment contains 228 (57 × 4 channels) features. The SVM classifier was trained and validated using LOSO cross-validation. The hyperparameters of the SVM with radial basis function kernel are set via 6-fold cross validation on the training data.

3. Results

For the baseline, interrater agreement between human experts was assessed for the quiet sleep annotations in the full training set. The overall agreement was very high (Cohen Kappa = 0.71, 95 %CI: 0.66–0.73; Fig. 2A) and represents the upper limit of achievable classifier performance.

3.1. Algorithm performance on the training dataset

The performance of the algorithm in detecting quiet sleep was assessed with LOSO cross-validation, using single channel detection (applied to F3-P3, F4-P4, F3-F4, and P3-P4) and the mild temporal smoothing. As shown in Fig. 2B, the algorithm performance against each of the experts is overlapping with the interrater agreement between the experts (Fig. 2A), and the classifier-expert agreement was very strong as well (Cohen kappa = 0.76, 95 %CI:0.73–0.79; Supplementary Table S2). Comparison of EEG
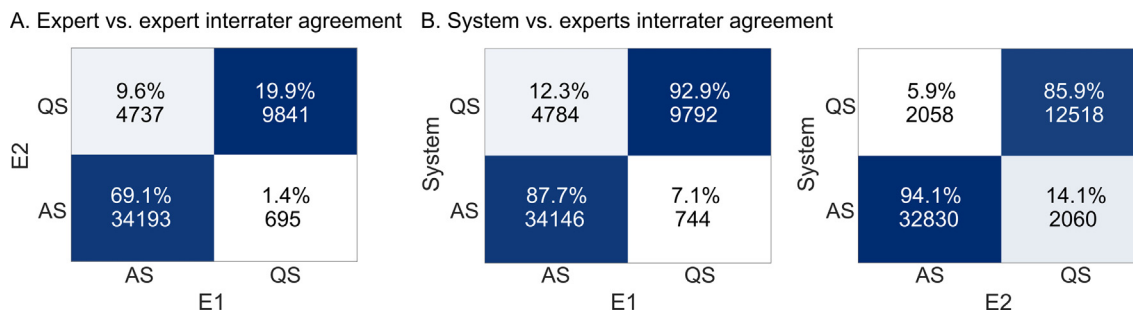
S. Montazeri Moghadam, Päivi Nevalainen, N.J. Stevenson et al.

Clinical Neurophysiology 143 (2022) 75–83

A. Expert vs. expert interrater agreement   B. System vs. experts interrater agreement

Fig. 2. Confusion matrices of agreement of sleep state classification. **A.** between experts. The percentages (and corresponding colours) are with respect to the total number of epochs. **B.** between the proposed classifier on the y-axis and each expert on the x-axis. The percentages (and corresponding colours) denote the recall value of each category. AS: active sleep, QS: quiet sleep.

derivations showed that the results generalize well across channels (Supplementary Table S1). To further explore the classifier performance vs ambiguity in the EEG, we compared classifier performance levels also against each expert separately, against all epochs, as well as against the epochs with full consensus between experts (disputed epochs (11 % of all epochs) are not considered in this case). Finally, the results were compared to the feature-based reference classifier. The results are summarized in Table 1 and Supplementary Table S2. This comparison indicated expectedly (cf. (Airaksinen et al., 2020, Moghadam et al., 2021)) that classification is a bit more accurate when considering the consensus epochs only. Notably, comparison of accuracies in individual neonates showed that detection was a clinically acceptable range (>80 %) in all neonates. Performance of the feature-based classifier was generally high as well but the accuracy in individual neonates could fall considerably lower challenging its utility in the clinical workup.

Finally, we estimated how well a straightforward amplitude envelope or its standard deviation computed from the EEG signal could correlate with the SST outputs. As shown in Supplementary Fig. S3, both the amplitude envelope itself and its standard deviation are noisy signals for the purpose of sleep classification, and correspondingly, they only show a week correlation with the SST output.

3.2. Validation with an independent dataset

To further validate the performance and generalizability of our classifier, we tested it using a PSG dataset, which uses different EEG derivations, a full PSG recording, and has been scored according to the American academy of sleep medicine (AASM) consensus guidelines rather than the EEG background alone (Fig. 3). We estimated algorithm's performance in distinguishing N3 (taken to represent QS) from the rest, i.e. wake/REM/N1 (taken to represent active sleep). The overall performance in terms of accuracy was 81 % (63–100; the classifier that performed the best in the LOSO cross-validation was used) in N3 versus other sleep states which

approaches levels of the inter-rater agreement reported for PSG scoring (Satomaa et al., 2016) (Fig. 3A).

We also compared the classifier output against all four PSG-derived sleep states. This analysis (Fig. 3B) shows that AS in our detector output most closely represents wake and REM while QS represents N3. There is a confusion between AS and QS for N1; this is expected given the transitory, dynamic nature of N1 state between wake and deeper sleep (Grigg-Damberger et al., 2007).

3.3. Visualization with Sleep State Trend (SST)

In order to be clinically useful, the algorithmic outputs need to be visualized in the bedside monitors. Visualization of sleep state detections is most natural using trends akin to those that are already used in the aEEG displays (Thoresen et al., 2010), seizure detections (Ansari et al., 2019, Stevenson et al., 2019, Stevenson and Vanhatalo, 2018, Tapani et al., 2019), or the vital signs monitors. To this end, we constructed an intuitive SST trend that depicts the weighted average of the probability outputs from the sleep state classifier for every 1-min EEG epoch. A comparable visualization of EEG background (Moghadam et al., 2021) was well received by a representative collection of clinicians. An example of SST is shown in Fig. 4 for a 24-hour recording in a neonate from our training dataset. The aEEG trend and the human expert annotations based on the raw EEG are shown for comparison. Sleep state fluctuations are challenging to observe in sections of aEEG trend, especially when it becomes contaminated by the NICU-typical artifacts such as cardiac activity, movements or high-frequency respiration, or when longer time epochs need to be visualized in the same display. The SST is, however, able to depict the sleep state relevant changes with clarity.

4. Discussion

Here we show an accurate detection of QS epochs using a CNN-based classifier on single EEG channel data from newborn infants near term equivalent age. The accuracy of this novel classifier is comparable in EEG channels that are typically used in the long
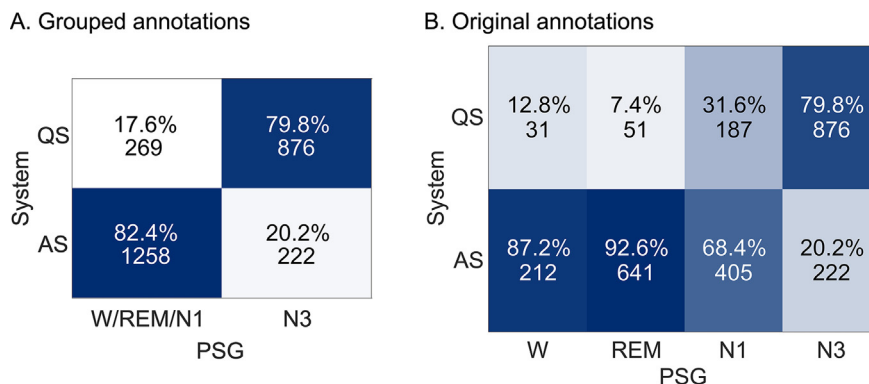
Table 1
Performance comparison between the reference feature-based algorithm and proposed classifier tested on consensus and all the labels. Since the feature-based classifier is based on multi-channel EEG, number of used epochs for its validation is about one-fourth of number of consensus epochs. PT: feature-based algorithm trained on EEG from preterm (Koolen et al., 2017). T: feature-based algorithm re-trained on EEG from term infants.

	# Epochs	Accuracy [%]	F1-score [%]	Precision [%]	Kappa
Feature-based (PT)	10,005	73 (55–86)	47 (40–56)	50 (40–57)	0.2 (0.1–0.3)
Feature-based (T)	10,005	87 (70–98)	77 (50–95)	76 (50–92)	0.5 (0.3–0.9)
Proposed CNN					
Consensus epochs	44,034	95 (83–99)	93 (80–99)	96 (89–99)	0.9 (0.6–1)
All epochs	98,932	90 (82–95)	88 (79–92)	90 (81–95)	0.8 (0.6–0.8)

S. Montazeri Moghadam, Päivi Nevalainen, N.J. Stevenson et al.

Clinical Neurophysiology 143 (2022) 75–83

Fig. 3. Confusion matrices of agreement of sleep state classification between the proposed classifier on the y-axis and PSG on the x-axis. Numbers are normalized such that each column sums up to one. AS: active sleep, QS: quiet sleep, W: wakefulness, REM: rapid eye movement, N1 & N3: non-REM stages 1 and 3, respectively.

term aEEG monitoring in NICUs. Importantly, the classifier performance generalizes well to an unseen EEG data consisting of other EEG derivations taken from an independent validation dataset of PSG recordings; this also provides a gold standard benchmark. Finally, we show that a classifier output of this kind can be visualized in an intuitive manner as a SST trend, which is a continuous display of QS probability and its confidence for quality assessment; importantly, this allows a continuous visual display that can be directly implemented into the bedside (a)EEG monitors. The main novelty of this work is in providing the full pipeline, a complete end-to-end solution from the raw signal to a validated classifier and its implementation into bedside monitors. Moreover, we offer this analytic pipeline openly for any future research use to expedite its take-up into clinical research.

Several prior studies have described automated methods for sleep state detection in the neonatal EEG signals (Ansari et al., 2022, 2020, 2018; Dereymaeker et al., 2017b; Fraiwan et al., 2011; Fraiwan and Alkhodari, 2020; Ghimatgar et al., 2020; Hsu et al., 2013; Koolen et al., 2017; Pillay et al., 2018; Piryatinska et al., 2009). The earlier solutions were commonly based on computational features extracted from multiple EEG channels and then combined in e.g. SVM-classifiers (Dereymaeker et al., 2017b; Fraiwan et al., 2011; Ghimatgar et al., 2020; Hsu et al., 2013; Koolen et al., 2017; Pillay et al., 2018; Piryatinska et al., 2009). Their performance has largely depended on averaging over longer time epochs and combining information from multiple channels in order to show a clinically reasonable accuracy. The need for multiple EEG signals has been a bottleneck in solutions for neonatal EEG monitoring that typically work on one to three EEG channels. More recently, CNN approaches have been developed to provide end-to-end solutions without heuristic feature engineering; methods that have been shown to hold promise in many applications of neonatal EEG classification (Ansari et al., 2020). Previous attempts with CNN algorithms for QS detection have shown that they outperform feature-based methods, which was confirmed in our work as well.

A direct comparison between different classifiers, such as our CNN and feature-based methods, is challenged by their different time resolutions. For instance, our CNN model was trained for 1-min epochs while the Koolen's classifier was initially trained for 2.5-min epochs (Koolen et al., 2017). A shorter epoch length may lead to higher uncertainty in classification, and it is commonly tapered by smoothing in the post-processing phase. For the ultimate implementation in bedside monitors, however, the desired epoch length is defined by the length of sleep wake cycles (SWC); in the neonate, SWC occurs in the scale of tens of minutes (Curzi-Dascalova et al., 1988, Osredkar et al., 2005), hence the epoch lengths between 30 s and 10 min used in the classifier constructs are all acceptable to provide measures of cycling (Stevenson

et al., 2014) or trends of sleep state fluctuations (Koolen et al., 2017).

Our findings are fully consistent with prior literature showing that QS detection is possible from the neonatal EEG, however we extend previous studies by developing a complete end-to-end solution that takes in single EEG channel data and provides a clinically useful SST visualization. Unlike the prior studies with multi-channel data, here we used only single EEG channels, which makes the findings useful as a bedside trend in the (a) EEG monitors. Our findings show an accuracy that compares well with the human interrater agreement, which sets the upper boundary of achievable classifier performance in tasks where the target label is based on subjective, visual interpretation. There are many possible reasons for the improved performance of our novel quiet sleep classifier compared to prior works. For instance, we used a large dataset and annotated the EEG signals in our training data without pre-fixed epochs, which was used in prior studies following the conventional and physiologically imperfect practice in sleep medicine. Larger datasets permit training of larger CNN architectures which increases the potential of the CNN to generalize to unseen data (cf. (Bubeck and Sellke, 2021)). We also incorporated annotations from multiple experts rather than consensus or single annotations as was done previously. Such inclusion of both experts' agreements and disagreements provides a more complex, but accurate, labelling of the data and has been shown to improve classifier performance (Airaksinen et al., 2020, Moghadam et al., 2021). We show here that the present algorithm generalizes well between derivations in the training dataset, as well as to completely different derivations, recording systems and sleep scoring systems used in our independent validation dataset.

In addition to describing the novel QS detector algorithm, we also present an intuitively interpretable visualization of the classifier output, SST, which allows direct implementation in the bedside EEG monitors. SST is not equivalent to the hypnogram that is generated by a human reviewer using international guidelines (Grigg-Damberger et al., 2007), but it allows a simple estimate of sleep cycling between AS and QS which may be clinically useful in many situations (Kidokoro et al., 2012). Unlike prior classifiers trained on healthy neonates, our classifier was trained on neonates that had recovered from serious conditions better reflecting the clinical situation we envision for the SST; tracking evolving SWC from imminent to mature patterns (see Fig. 4).

A future clinical implementation of our work needs to consider two limitations: First, the classifier only works for EEG recording from infants at near term age. As shown with our comparison to the reference classifier (Table 1), the sleep detectors may not generalize well between preterm and term age EEG records due to dramatic developmental changes that take place during this period
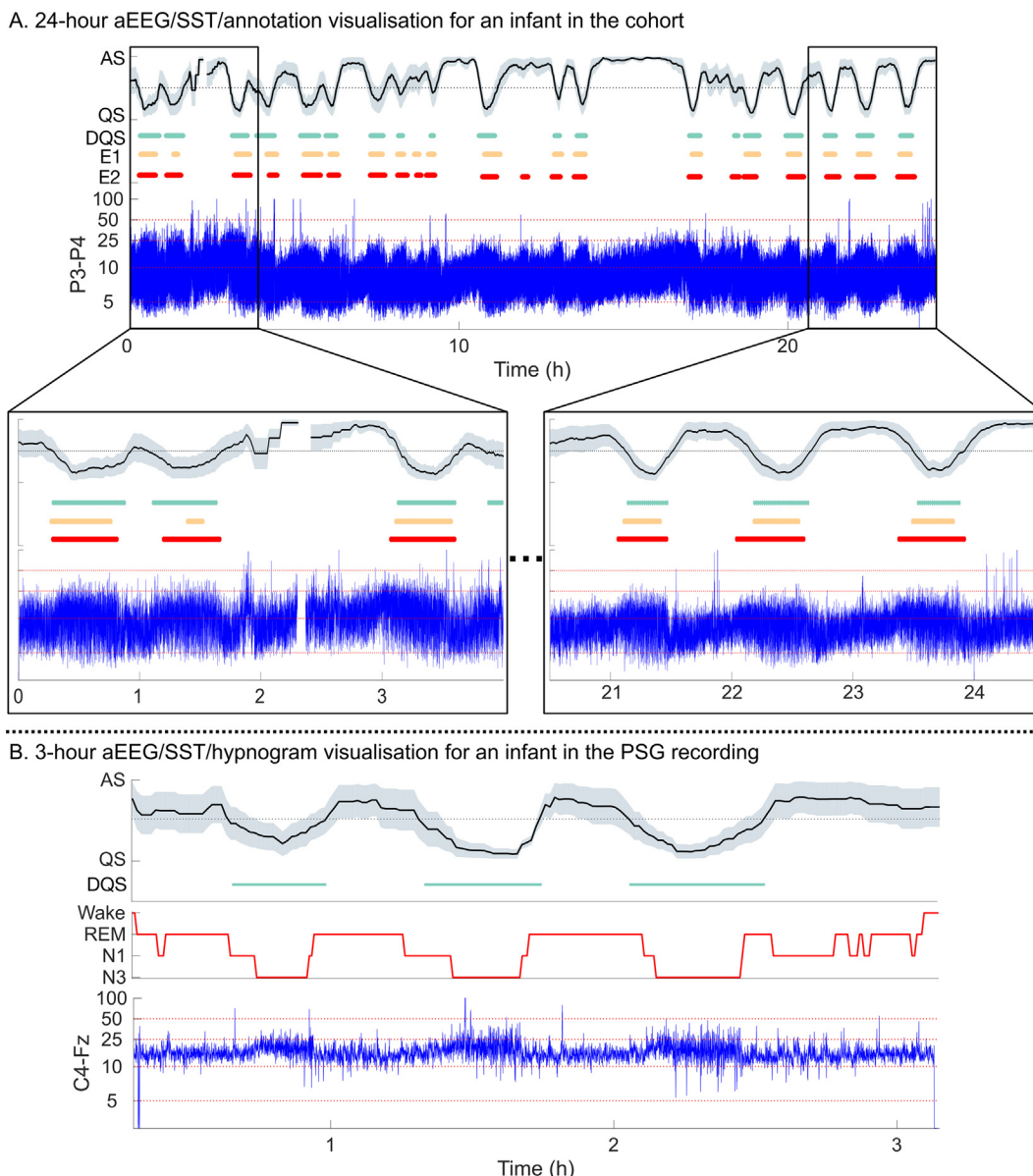
Fig. 4. Example SST results. The actual SST (black line) is complemented by depicting the uncertainty of the classifier (gray shadows) at each time point to provide the clinician with a quality index of classification. The uncertainty is quantified by the distribution of the probability outputs of the classifier. Using a fixed threshold (black dotted line in the SST) allows plotting a dichotomic detection of QS states (DQS) which are depicted with green lines under the SST. Expert annotations (E1 and E2) are shown for comparison. **A**. A full 24-hour signal from the typical P3-P4 derivation. Please note the clear distinction between sleep states in the SST output, with high agreement with the experts' annotations, and a clear change from a poorly organized ("imminent") to a well-organized ("mature") SWC during this time period. This is hard to recognize from the aEEG with 24 hours view, but it is relatively clear in the two 4-hour views of aEEG. The brief discontinuities in the SST output are due to removal of artefactual epochs in the preprocessing stage. **B**. Example of the SST output computed from about 3 hours of PSG recording. The hypnogram depicts the sleep states annotated according to the AASM criteria. Note the overall high agreement between algorithmic detection (SST and DQS) vs hypnogram. AS: active sleep, QS: quiet sleep.

(Bourel-Ponchel et al., 2021, Dereymaeker et al., 2017a, Stevenson et al., 2020, Vanhatalo and Kaila, 2006). Ideally, such an omnipotent classifier could be developed in a hierarchical fashion where the infant's age is first analyzed using the Functional Brain Age algorithm (FBA, (Stevenson et al., 2020)), followed by a sleep state detector that uses FBA as an input. Development of such flexible classifier would be essential to support wide-scale prospective studies on sleep-oriented care and the use of sleep as a functional biomarker across the wide range of conceptional ages present in the NICUs. Second, ambiguity in sleep states may be even more pronounced than in our dataset when monitoring neonates after recovery of cerebral injury, such as birth asphyxia (Thoresen et al., 2010). Evolution of the gradually emerging SWC is considered to be of key interest, however it also presents a conceptual

challenge to define sleep stages that emerge from a discontinuous overall background activity. These issues go beyond algorithms, and require characterization by the neonatal community before its faithful detection can be requested from computational algorithms.

A future validation effort that includes prospective data collection, clinically relevant contexts, and a consensus-based definition of conclusions that should be drawn from different means is needed. To make this possible, we present our classifier openly via our Babacloud server (www[dot]babacloud[dot]fi) to anyone interested in its exploitation in clinical trials. Future research is also needed to define the relative importance of detecting quiet sleep vs other vigilance states, and an accurate, specific measure of other sleep states will require corresponding detector development.

S. Montazeri Moghadam, Päivi Nevalainen, N.J. Stevenson et al.

Clinical Neurophysiology 143 (2022) 75–83

Detection of any sleep state is sufficient for the purpose of estimating sleep-wake cyclicity (Kidokoro et al., 2012, Stevenson et al., 2014). Since visual review of intermittent QS epochs in the aEEG trend has become the bedside method of choice, our SST trend allows a smooth transition between aEEG and SST interpretations.

When assessing the clinical or scientific utility of a novel method, it is important to consider the potential ambiguities in human annotations, in the data, as well as in the scoring systems; moreover, these should be considered with respect to the aimed implementation. Here, we aimed to develop a classifier that could provide a bedside index of SWC. To this end, it is necessary to have a reliable enough detection of at least one clear sleep state, i.e. QS in our case. Prior literature has shown that QS in the near term infants is mostly recognized as a mixture of tracé alternant and/or high voltage synchronous EEG pattern with an imperfect but good enough accuracy (Andre et al., 2010, Dereymaeker et al., 2017a), and conversely, tracé alternant pattern or its aEEG equivalent has become the hallmark of QS in the clinical work on SWC in the NICU monitoring (Kidokoro et al., 2012, Schwindt et al., 2015, Thoresen et al., 2010). We show that our QS detector works with an accuracy that is equivalent to the level agreement of the human experts visual interpretation in both the training and validation datasets. Hence, the difference between the algorithm's output and an individual human annotations may simply reflect ambiguity in the data itself as much as an error in the algorithm's output. This ambiguity in sleep states becomes even more apparent when considering the spectrum of elusive states that characterize newborn brain recovery from critical illness, such as birth asphyxia (Thoresen et al., 2010). Therefore, the ultimate clinical utility of a novel method like SST should not be evaluated only by a strict comparison with the conventional laboratory methods; rather, one should estimate its perceived added value to the bedside clinician, a work that needs future prospective trials in multicenter settings (Pavel et al., 2020).

Funding

Conflict of interest

Authors declare no conflict of interest.

Author contributions

SM and SV conceived the experimental design. PN and SV performed the visual annotation. SM developed the CNN classifier designs, implemented it and analyzed the EEG data. NS developed the original version of the feature-based reference classifier. SM and SV prepared the manuscript and figures. All authors carefully reviewed, commented, and finally approved the article.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.clinph.2022.08.022.

References

Airaksinen M, Rasanen O, Ilen E, Hayrinen T, Kivi A, Marchi V, et al. Automatic Posture and Movement Tracking of Infants with Wearable Movement Sensors. Sci Rep 2020;10(1):169. https://doi.org/10.1038/s41598-019-56862-5.

Allen KA. Promoting and protecting infant sleep. Adv Neonatal Care 2012;12 (5):288–91. https://doi.org/10.1097/ANC.0b013e3182653899.

André M, Lamblin MD, d'Allest AM, Curzi-Dascalova L, Moussalli-Salefranque FT, Tich SN, et al. Electroencephalography in premature and full-term infants. Developmental features and glossary. Neurophysiol Clin 2010;40(2):59–124. https://doi.org/10.1016/j.neucli.2010.02.002.

Ansari AH, Cherian PJ, Caicedo A, Naulaers G, De Vos M, Van Huffel S. Neonatal seizure detection using deep convolutional neural networks. Int J Neural Syst 2019;29(04):1850011.

Ansari AH, De Wel O, Lavanga M, Caicedo A, Dereymaeker A, Jansen K, et al. Quiet sleep detection in preterm infants using deep convolutional neural networks. J Neural Eng 2018;15(6). https://doi.org/10.1088/1741-2552/aadc1f 066006.

Ansari AH, De Wel O, Pillay K, Dereymaeker A, Jansen K, Van Huffel S, et al. A convolutional neural network outperforming state-of-the-art sleep staging algorithms for both preterm and term infants. J Neural Eng 2020;17(1). https://doi.org/10.1088/1741-2552/ab5469 016028.

Ansari AH, Pillay K, Dereymaeker A, Jansen K, Van Huffel S, Naulaers G, et al. A Deep Shared Multi-Scale Inception Network Enables Accurate Neonatal Quiet Sleep Detection With Limited EEG Channels. IEEE J Biomed Health Inform 2022;26 (3):1023–33. https://doi.org/10.1109/JBHI.2021.3101117.

Arditi-Babchuk H, Feldman R, Eidelman AI. Rapid eye movement (REM) in premature neonates and developmental outcome at 6 months. Infant Behav Dev 2009;32(1):27–32. https://doi.org/10.1016/j.infbeh.2008.09.001.

Bishop CM. Neural networks for pattern recognition. Oxford University Press; 1995.

Bourel-Ponchel E, Hasaerts D, Challamel MJ, Lamblin MD. Behavioral-state development and sleep-state differentiation during early ontogenesis. Neurophysiol Clin 2021;51 (1):89–98. https://doi.org/10.1016/j.neucli.2020.10.003.

Bubeck S, Sellke M. A universal law of robustness via isoperimetry. Adv Neural Inf Process Syst 2021;34.

Cherian PJ, Swarte RM, Visser GH. Technical standards for recording and interpretation of neonatal electroencephalogram in clinical practice. Ann Indian Acad Neurol 2009;12(1):58–70. https://doi.org/10.4103/0972-2327.48869.

Curzi-Dascalova L, Peirano P, Morel-Kahn F. Development of sleep states in normal premature and full-term newborns. Dev Psychobiol 1988;21(5):431–44. https://doi.org/10.1002/dev.420210503.

Dereymaeker A, Pillay K, Vervisch J, De Vos M, Van Huffel S, Jansen K, et al. Review of sleep-EEG in preterm and term neonates. Early Hum Dev 2017a;113:87–103. https://doi.org/10.1016/j.earlhumdev.2017.07.003.

Dereymaeker A, Pillay K, Vervisch J, Van Huffel S, Naulaers G, Jansen K, et al. An Automated Quiet Sleep Detection Approach in Preterm Infants as a Gateway to Assess Brain Maturation. Int J Neural Syst 2017b;27(6):1750023. https://doi.org/10.1142/S012906571750023X.

Ednick M, Cohen AP, McPhail GL, Beebe D, Simakajornboon N, Amin RS. A review of the effects of sleep during the first year of life on cognitive, psychomotor, and temperament development. Sleep 2009;32(11):1449–58. https://doi.org/10.1093/sleep/32.11.1449.

Fraiwan L, Alkhodari M. Neonatal sleep stage identification using long short-term memory learning system. Med Biol Eng Comput 2020;58(6):1383–91. https://doi.org/10.1007/s11517-020-02169-x.

Fraiwan L, Lweesy K, Khasawneh N, Fraiwan M, Wenz H, Dickhaus H. Time frequency analysis for automated sleep stage identification in fullterm and preterm neonates. J Med Syst 2011;35(4):693–702. https://doi.org/10.1007/s10916-009-9406-2.

Ghimatgar H, Kazemi K, Helfroush MS, Pillay K, Dereymaker A, Jansen K, et al. Neonatal EEG sleep stage classification based on deep learning and HMM. J Neural Eng 2020;17(3). https://doi.org/10.1088/1741-2552/ab965a 036031.

Graven S. Sleep and brain development. Clin Perinatol 2006;33(3):693–706. https://doi.org/10.1016/j.clp.2006.06.009.

Grigg-Damberger M, Gozal D, Marcus CL, Quan SF, Rosen CL, Chervin RD, et al. The visual scoring of sleep and arousal in infants and children. J Clin Sleep Med 2007;3(2):201–40.

He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proc IEEE Int Conf Comput Vis 2015:1026–34.

Hsu Y, Yang Y, Wang J, Hsu C. Automatic sleep stage recurrent neural classifier using energy features of EEG signals. Neurocomputing 2013;104:105–14. https://doi.org/10.1016/j.neucom.2012.11.003.

Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Int conf mach learn (ICML): PMLR; 2015. p. 448–56.

Kidokoro H, Inder T, Okumura A, Watanabe K. What does cyclicity on amplitude-integrated EEG mean? J Perinatol 2012;32(8):565–9.

Koolen N, Oberdorfer L, Rona Z, Giordano V, Werther T, Klebermass-Schrehof K, et al. Automated classification of neonatal sleep states using EEG. Clin Neurophysiol 2017;128(6):1100–8. https://doi.org/10.1016/j.clinph.2017.02.025.

Lam P, Hiscock H, Wake M. Outcomes of infant sleep problems: a longitudinal study of sleep, behavior, and maternal well-being. Pediatrics 2003;111(3):e203–7. https://doi.org/10.1542/peds.111.3.e203.

S. Montazeri Moghadam, Päivi Nevalainen, N.J. Stevenson et al.

Clinical Neurophysiology 143 (2022) 75–83

Marks GA, Shaffery JP, Oksenberg A, Speciale SG, Roffwarg HP. A functional role for REM sleep in brain maturation. Behav Brain Res 1995;69(1–2):1–11. https://doi.org/10.1016/0166-4328(95)00018-o.

Meder U, Cseko AJ, Szakacs L, Balogh CD, Szakmar E, Andorka C, et al. Longitudinal Analysis of Amplitude-Integrated Electroencephalography for Outcome Prediction in Hypoxic-Ischemic Encephalopathy. J Pediatr 2022;246:19–25.e5. https://doi.org/10.1016/j.jpeds.2022.04.013.

Mirmiran M, Maas YG, Ariagno RL. Development of fetal and neonatal sleep and circadian rhythms. Sleep Med Rev 2003;7(4):321–34. https://doi.org/10.1053/smrv.2002.0243.

Moghadam SM, Pinchefsky E, Tse I, Marchi V, Kohonen J, Kauppila M, et al. Building an Open Source Classifier for the Neonatal EEG Background: A Systematic Feature-Based Approach From Expert Scoring to Clinical Visualization. Front Hum Neurosci 2021;15. https://doi.org/10.3389/fnhum.2021.675154 675154.

Nevalainen P, Metsaranta M, Marchi V, Toiviainen-Salo S, Vanhatalo S, Lauronen L. Towards multimodal brain monitoring in asphyxiated newborns with amplitude-integrated EEG and simultaneous somatosensory evoked potentials. Early Hum Dev 2021;153. https://doi.org/10.1016/j.earlhumdev.2020.105287 105287.

Nevalainen P, Metsaranta M, Toiviainen-Salo S, Lonnqvist T, Vanhatalo S, Lauronen L. Bedside neurophysiological tests can identify neonates with stroke leading to cerebral palsy. Clin Neurophysiol 2019;130(5):759–66. https://doi.org/10.1016/j.clinph.2019.02.017.

Osredkar D, Toet MC, van Rooij LG, van Huffelen AC, Groenendaal F, de Vries LS. Sleep-wake cycling on amplitude-integrated electroencephalography in term newborns with hypoxic-ischemic encephalopathy. Pediatrics 2005;115(2):327–32. https://doi.org/10.1542/peds.2004-0863.

Pavel AM, Rennie JM, de Vries LS, Blennow M, Foran A, Shah DK, et al. A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial. Lancet Child Adolesc Health 2020;4(10):740–9. https://doi.org/10.1016/S2352-4642(20)30239-X.

Phan H, Andreotti F, Cooray N, Chen OY, De Vos M. Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification. IEEE Trans Biomed Eng 2019a;66(5):1285–96. https://doi.org/10.1109/TBME.2018.2872652.

Phan H, Andreotti F, Cooray N, Chen OY, De Vos M. SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging. IEEE Trans Neural Syst Rehabil Eng 2019b;27(3):400–10. https://doi.org/10.1109/TNSRE.2019.2896659.

Pillay K, Dereymaeker A, Jansen K, Naulaers G, Van Huffel S, De Vos M. Automated EEG sleep staging in the term-age baby using a generative modelling approach. J Neural Eng 2018;15(3). https://doi.org/10.1088/1741-2552/aaab73 036004.

Piryatinska A, Terdik G, Woyczynski WA, Loparo KA, Scher MS, Zlotnik A. Automated detection of neonate EEG sleep stages. Comput Methods Programs Biomed 2009;95(1):31–46. https://doi.org/10.1016/j.cmpb.2009.01.006.

Roffwarg HP, Muzio JN, Dement WC. Ontogenetic development of the human sleep-dream cycle. Science 1966;152(3722):604–19. https://doi.org/10.1126/science.152.3722.604.

Satomaa AL, Saarenpaa-Heikkila O, Paavonen EJ, Himanen SL. The adapted American Academy of Sleep Medicine sleep scoring criteria in one month old infants: A

means to improve comparability? Clin Neurophysiol 2016;127(2):1410–8. https://doi.org/10.1016/j.clinph.2015.08.013.

Scher MS, Steppe DA, Banks DL. Prediction of lower developmental performances of healthy neonates by neonatal EEG-sleep measures. Pediatr Neurol 1996;14(2):137–44. https://doi.org/10.1016/0887-8994(96)00013-6.

Schwindt E, Thaller C, Czaba-Hnizdo C, Giordano V, Olischar M, Waldhoer T, et al. Being born small for gestational age influences amplitude-integrated electroencephalography and later outcome in preterm infants. Neonatology 2015;108(2):81–7.

Sors A, Bonnet S, Mirek S, Vercueil L, Payen J-F. A convolutional neural network for sleep stage scoring from raw single-channel EEG. Biomed Signal Process Control 2018;42:107–14.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res (JMLR) 2014;15(1):1929–58.

Stevenson NJ, Oberdorfer L, Tataranno ML, Breakspear M, Colditz PB, de Vries LS, et al. Automated cot-side tracking of functional brain age in preterm infants. Ann Clin Transl Neurol 2020;7(6):891–902. https://doi.org/10.1002/acn3.51043.

Stevenson NJ, Palmu K, Wikstrom S, Hellstrom-Westas L, Vanhatalo S. Measuring brain activity cycling (BAC) in long term EEG monitoring of preterm babies. Physiol Meas 2014;35(7):1493–508. https://doi.org/10.1088/0967-3334/35/7/1493.

Stevenson NJ, Tapani K, Lauronen L, Vanhatalo S. A dataset of neonatal EEG recordings with seizure annotations. Sci Data 2019;6. https://doi.org/10.1038/sdata.2019.39 190039.

Stevenson NJ, Vanhatalo S. Designing a trial for neonatal seizure treatment. Semin Fetal Neonatal Med 2018;23(3):213–7. https://doi.org/10.1016/j.siny.2018.02.005.

Tapani KT, Vanhatalo S, Stevenson NJ. Time-Varying EEG Correlations Improve Automated Neonatal Seizure Detection. Int J Neural Syst 2019;29(4):1850030. https://doi.org/10.1142/S0129065718500302.

Thoresen M, Hellstrom-Westas L, Liu X, de Vries LS. Effect of hypothermia on amplitude-integrated electroencephalogram in infants with asphyxia. Pediatrics 2010;126(1):e131–9. https://doi.org/10.1542/peds.2009-2938.

Touchette E, Petit D, Seguin JR, Boivin M, Tremblay RE, Montplaisir JY. Associations between sleep duration patterns and behavioral/cognitive functioning at school entry. Sleep 2007;30(9):1213–9. https://doi.org/10.1093/sleep/30.9.1213.

van den Hoogen A, Teunis CJ, Shellhaas RA, Pillen S, Benders M, Dudink J. How to improve sleep in a neonatal intensive care unit: a systematic review. Early Hum Dev 2017;113:78–86.

Vanhatalo S, Kaila K. Development of neonatal EEG activity: from phenomenology to physiology. Semin Fetal Neonatal Med 2006;11(6):471–8. https://doi.org/10.1016/j.siny.2006.07.008.

Zhang J, Wu Y. Automatic sleep stage classification of single-channel EEG by using complex-valued convolutional neural network. Biomed Tech (Berl) 2018;63(2):177–90. https://doi.org/10.1515/bmt-2016-0156.

Zhang J, Yao R, Ge W, Gao J. Orthogonal convolutional neural networks for automatic sleep stage classification based on single-channel EEG. Comput Methods Programs Biomed 2020;183. https://doi.org/10.1016/j.cmpb.2019.105089 105089.