Journal of Neural Engineering

PAPER

Neonatal EEG sleep stage classification based on deep learning and HMM

To cite this article: Hojat Ghimatgar *et al* 2020 *J. Neural Eng.* **17** 036031

View the article online for updates and enhancements.

PITT | SWANSON ENGINEERING
BIOENGINEERING

The Department of Bioengineering at the University of Pittsburgh Swanson School of Engineering invites applications from accomplished individuals with a PhD or equivalent degree in bioengineering, biomedical engineering, or closely related disciplines for an open-rank, tenured/tenure-stream faculty position. We wish to recruit an individual with strong research accomplishments in Translational Bioengineering (i.e., leveraging basic science and engineering knowledge to develop innovative, translatable solutions impacting clinical practice and healthcare), with preference given to research focus on neuro-technologies, imaging, cardiovascular devices, and biomimetic and biorobotic design. It is expected that this individual will complement our current strengths in biomechanics, bioimaging, molecular, cellular, and systems engineering, medical product engineering, neural engineering, and tissue engineering and regenerative medicine. In addition, candidates must be committed to contributing to high quality education of a diverse student body at both the undergraduate and graduate levels.

To ensure full consideration, applications must be received by June 30, 2019. However, applications will be reviewed as they are received. Early submission is highly encouraged.

CLICK HERE FOR FURTHER DETAILS

This content was downloaded from IP address 134.58.253.30 on 24/09/2020 at 08:22

Journal of Neural Engineering



RECEIVED
17 January 2020

REVISED
8 May 2020

ACCEPTED FOR PUBLICATION
26 May 2020

PUBLISHED
25 June 2020

PAPER

Neonatal EEG sleep stage classification based on deep learning and HMM

Hojat Ghimatgar[1,2], Kamran Kazemi[1], Mohammad Sadegh Helfroush[1], Kirubin Pillay[3,4], Anneleen Dereymaker[5], Katrien Jansen[5,6], Maarten De Vos[7] and Ardalan Aarabi[8,9]

[1] Department of Electrical and Electronics Engineering, Shiraz University of Technology, Shiraz, Iran
[2] Department of Electrical Engineering, Persian Gulf University, Bushehr, Iran
[3] Institute of Biomedical Engineering (IBME), Department of Engineering Science, University of Oxford, Oxford, United Kingdom
[4] Department of Paediatrics, John Radcliffe Hospital, University of Oxford, Oxford, United Kingdom
[5] Department of Development and Regeneration, University Hospitals Leuven, Neonatal Intensive Care Unit, KU Leuven (University of Leuven), Leuven, Belgium
[6] Department of Development and Regeneration, University Hospitals Leuven, Child Neurology, University of Leuven (KU Leuven), Leuven, Belgium
[7] ESAT – Stadius KU Leuven, Leuven, Belgium
[8] Laboratory of Functional Neuroscience and Pathologies (LNFP), University Research Center (CURS), University Hospital, Amiens, France
[9] Faculty of Medicine, University of Picardie Jules Verne, Amiens, France

E-mail: ardalan.aarabi@u-picardie.fr

Keywords: EEG, sleep stage classification, feature selection, channel selection, deep learning, HMM, neonates

Supplementary material for this article is available online

Abstract

Objective. Automatic sleep stage scoring is of great importance for investigating sleep architecture during infancy. In this work, we introduce a novel multichannel approach based on deep learning networks and hidden Markov models (HMM) to improve the accuracy of sleep stage classification in term neonates. *Approach*. The classification performance was evaluated on quiet sleep (QS) and active sleep (AS) stages, each with two sub-states, using multichannel EEG data recorded from sixteen neonates with postmenstrual age of 38–40 weeks. A comprehensive set of linear and nonlinear features were extracted from thirty-second EEG segments. The feature space dimensionality was then reduced by using an evolutionary feature selection method called MGCACO (Modified Graph Clustering Ant Colony Optimization) based on the relevance and redundancy analysis. A bi-directional long-short time memory (BiLSTM) network was trained for sleep stage classification. The number of channels was optimized using the sequential forward selection method to reduce the spatial space. Finally, an HMM-based postprocessing stage was used to reduce false positives by incorporating the knowledge of transition probabilities between stages into the classification process. The method performance was evaluated using the K-fold (KFCV) and leave-one-out cross-validation (LOOCV) strategies. *Main results*. Using six-bipolar channels, our method achieved a mean kappa and an overall accuracy of 0.71–0.76 and 78.9%–82.4% using the KFCV and LOOCV strategies, respectively. *Significance*. The presented automatic sleep stage scoring method can be used to study the neurodevelopmental process and to diagnose brain abnormalities in term neonates.

1. Introduction

Human newborns spend two-thirds of their lives sleeping. The quality of sleep plays an important role in infant neurodevelopment especially in the first year of life. In infants, it is very difficult to study brain functions through cognitive examinations. This is why sleep studies can be very useful to investigate the neurodevelopmental process and brain abnormalities during the infancy period by exploring sleep structure.

In sleep studies, electroencephalography (EEG) is routinely used as an efficient tool to systematically investigate sleep structure and stages. Each sleep stage is identified with a set of specific physiological markers accompanied by characteristics changes in

© 2020 IOP Publishing Ltd

spectral features, which are used to explore the sleep architecture including the timing and organization of sleep stages over the course of sleep.

In adults, sleep stages are classified into wake, rapid eye movement (REM) and non-REM (N1, N2, and N3) according to the American Academy of Sleep Medicine (AASM) scoring manual [1]. Compared to adults, neonates show markedly low homogeneity and instability in EEG during wake or sleep [2]. In neonates, two major sleep states are clearly distinguished, the Quiet Sleep (QS equivalent to non-REM sleep in adults) and Active Sleep (AS equivalent to REM sleep), with in-between indeterminate or transitional epochs exhibiting non-EEG characteristics of AS and EEG characteristics of QS [2, 3]. The transitional epochs are often observed at sleep onset (arousal state) and between REM and non-REM sleep [3–5]. Overall, term neonates spend 50%–60% of their sleep cycles in AS, 30%–40% in QS, and 10%–15% in the transitional state [4], with sleep-wake and sleep-only cycles lasting typically 3–4 h and 40–70 min, respectively [4].

At term-age, two sub-states AS_I and AS_{II} appear within AS (REM sleep) with distinct temporal and spectral features. AS_I is characterized by mixed frequencies and variable amplitudes and AS_{II} with Low-Voltage Irregular (LVI) patterns [2]. In full-term neonates, in addition to EEG, other physiological and behavioral variables are required to identify wakefulness and REM sleep, both exhibiting LVI or continuous mixed frequencies patterns [3–5]. At full-term age, QS also shows two sub-states, Tracé Alternant (TA) and High Voltage Slow-wave (HVS) [6]. As the most abundant activity, TA is characterized by suppression burst periods with increased amplitudes. The HVS, however, exhibits EEG activities of higher amplitude and lower frequency in comparison with AS [7].

Visual sleep stage scoring is performed by trained neurologists, a time-consuming and costly procedure that is highly prone to human errors depending on experts' experiences. Automatic sleep staging can achieve high accuracy and efficiently reduce scoring time.

In adults, many approaches relying on various classifiers have been developed for sleep staging using biosignals (EEG, electrocardiogram (ECG), electromyogram (EMG), electrooculogram (EOG) and respiratory signals) [8–16]. The use of multichannel and multiple biosignals for sleep staging requires a complicated long-time procedure, which limits the subject's movements and may affect patients' normal sleep [13, 17]. Moreover, in adults due to the functional symmetry between bilateral brain regions, single-channel processing can provide relatively high accuracies with much less computational complexities in comparison with sleep staging algorithms based on multichannel data [12, 17].

In neonates, a few automatic sleep staging methods have been introduced based on a variety of features and classifiers with a limited degree of success [18–23] due to several technical problems. First, the majority of these methods have employed features commonly used in sleep studies in adults [12, 24], including time and frequency domain as well as nonlinear and complexity features [2, 25]. Due to the brain immaturity in neonates, their sleep EEG signals show nonstationary and nonlinear characteristics that make automatic sleep stage classification difficult and result in low accuracy rates of automated tools. In very few studies, neonate-specific features relying on *a priori* knowledge about EEG characteristics have been used to improve prediction performance for sleep staging in neonates [20, 21]. In this context, feature selection or dimensionality reduction, an efficient way to improve the classification performance and decrease the computational complexity, has gained little attention in sleep studies in neonates [2, 20, 25, 26]. Second, compared to the wide range of machine learning-based methods used for sleep stage classification in adults [12], very few efforts have been made towards developing efficient tools based on learning tools with complicated architectures including deep learning networks, which have shown good generalization abilities for sleep staging especially using class imbalanced data [23, 27–29]. Third, very few studies have focused on the benefit of channel selection methods, which could efficiently decrease the computational complexity and improve classification accuracy [6]. It is of note that due to asymmetry between brain regions in neonates, all channels do not contribute equally to the sleep staging classification process. This is why the performance of single-channel methods provides relatively lower accuracy rates compared to multi-channel methods developed for sleep staging in neonates. Finally, due to the cyclic nature of sleep stages especially between non-REM/REM, incorporating the temporal structure of sleep cycles into the classification process would improve the prediction performance. In adults, a few studies have taken into account the temporal organization of sleep stages to reduce false positives of the sleep stage classification [12, 30, 31]. In neonates, very few studies have used temporal sleep stage classification methods by modeling the sequential nature of sleep stages using hidden Markov models (HMM) or some expert rules [2, 23]. The sleep stage transition rules embedded in predictive models can efficiently improve the classification performance by incorporating information from the neighboring time segments used mainly to rule out infrequent stage transitions.

The main purpose of the present paper, motivated by our approach proposed for EEG sleep stage scoring in adults [12], was to develop an automated multichannel EEG sleep stage classification method

in full-term neonates. To improve the prediction performance and reduce the computational complexity, the method relied on a four-stage processing pipeline—feature extraction and selection, channel selection, classification, and postprocessing. In the first stage after preprocessing, a comprehensive set of time, frequency, and nonlinear features widely used in sleep studies in adults and neonates [2, 12] was extracted from EEG segments. An efficient feature selection method called MGCACO (Modified Graph Clustering Ant Colony Optimization) [26] was then used to select optimal feature sets based on the relevance and redundancy analysis. A Bidirectional Long Short-Term Memory (BiLSTM) network was trained using the optimal feature sets and used for sleep stage classification. To improve the classification performance, a channel selection method was used to optimize the spatial sequence of EEG channels. Finally, in the postprocessing stage, an HMM was employed to incorporate the temporal organization of sleep cycles into the sleep stage classification procedure. The performance of the method was evaluated using the KFCV and LOOCV strategies on sleep EEG from full-term healthy neonates [2].

2. Methods

2.1. Experimental data

We used the EEG sleep data recorded from 16 full-term healthy neonates of 38–40 weeks postmenstrual age at the neonatal intensive care unit (NICU) of the University Hospitals, Leuven, Belgium for method development and evaluation. The data collection was approved by the ethics committee of the University Hospitals, Leuven [2]. In total, sixteen EEG sleep recordings with an average data length of 7 h 55 min (1 h 58 min–17 h 50 min) were collected from neonates selected based on a 'normal' developmental outcome [2]. The EEG data were recorded by using nine mono-polar EEG channels (Fp1, Fp2, T3, C3, Cz, C4, T4, O1 and O2) with a sampling frequency of 250 Hz according to the international 10–20 system of electrode placement. Table 1 lists the specification of the neonates and the data used for analysis.

The manual staging was visually performed by two EEG experts, who identified four sleep stages Active Sleep 1 (AS_I), Active Sleep 2 (AS_{II}, LVI), Quiet Sleep 1 (HVS) and Quiet sleep 2 (TA) as well as two other states defined as indeterminate (IS) and artifacts. We considered LVI as the state representing both Wake and AS_{II} due to their similarities in EEG patterns [3, 32]. In neonates specifically, EEG signals combined with other physiological and behavioral variables including EOG, EMG, vocalization (whimpering, crying) as well as the irregularity/regularity of respiration are required to identify Wake and AS_{II} as two separate states [3–5].

In accordance with the current recommendations and practical guidelines for sleep scoring in neonates [3–5], we partitioned the EEG signals into thirty-second epochs with an overlap of 5 sec between consecutive epochs to avoid losing transient changes at segment edges [2, 4]. The last segment in each stage was discarded from further analysis if its duration was less than 30 s. A single sleep stage label was assigned to each segment if both experts agreed on; otherwise, it was considered as dubious. For method development and evaluation, we used both monopolar (referential) and bipolar (longitudinal and transversal) montages as shown in figure 1 and table 1. All artifacts, indeterminate and dubious segments were discarded from further analysis.

Table 1. Specifications of the neonatal sleep EEG database used in this study.

Number of neonates	16
Age (postmenstrual age, PMA)	38–42 weeks
Sampling frequency (Hz)	250
Average data length per neonate (hours)	∼7.9
Total data length (hours)	∼126
Number of 30-second segments	9817
Number of segments per sleep stage	**Stage-segments (%)** AS_I—1915 (19.5%) HVS—1299 (13.2%) TA—2782 (28.3%) LVI—2492 (25.4%) IS—175 (1.7%) Dubious—1154 (11.8%)
EEG mono-polar channels	Fp1, Fp2, T3, C3, Cz, C4, T4, O1 and O2
EEG bipolar channels	Fp1-Fp2, T3-C3, C3-Cz, Cz-C4, C4-T4, T3-T4, C3-C4, O1-O2, Fp1-C3, C3-O1, Fp2-C4 and C4-O2

2.2. Proposed method

Figure 2 shows the block diagram of the multichannel sleep staging method based on the BiLSTM network and the HMM. In the training phase, the feature and channel selection, classifier training, and HMM optimization were performed using training sets. In the testing phase, the tool with the optimized configuration was evaluated on test sets.

2.2.1. Preprocessing

The EEG signals were first band-pass filtered using a fourth-order Butterworth filter within 0.5–30 Hz. Artifact-free EEG segments with definite sleep stage labels were used for method development based on inherent changes in characteristics of EEG signals in different sleep stages. For each subject *i*, artifact-free EEG segments were re-scaled on a channel-by-channel basis using Z-score normalization as follows:

$$x_{i,j}^{n} = \frac{x_{i,j} - m_{x_{i,j}}}{\sigma_{x_{i,j}}} \quad (1)$$

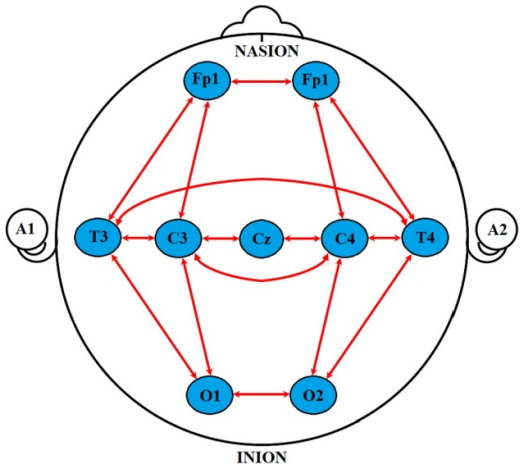

Figure 1. Monopolar (blue circles) and bipolar (red lines) channels used in this study.

where $x_{i,j}$ and $x^n_{i,j}$ were the EEG signal of the jth monopolar/bipolar channel before and after normalization, respectively, and $m_{x_{i,j}}$ and $\sigma_{x_{i,j}}$ were the mean and standard deviation of $x_{i,j}$.

2.2.2. Feature extraction and selection

To identify effective features best representing different sleep states in neonates, a comprehensive set of 168 features widely used in sleep studies in adults and neonates was extracted from each EEG segment [12, 13, 26, 33–35] as illustrated in figure 3. Among these features, 32 features were specifically chosen to characterize neonatal EEG signals.

The time-domain features consisted of a set of 52 morphological and statistical attributes extracted from each segment. The statistical features were used to characterize the signal distribution statistics, while the morphological features were used to characterize the overall attributes of EEG waveforms within the segment [33]. Seventy-three frequency-domain features were extracted from each segment within four frequency bands: δ (0.5–4 Hz), θ (4–8 Hz), α (8–13 Hz) and β (13–30 Hz), which are typically used as characteristic physiological bands of frequency in sleep research [12]. Among these features, twenty-nine frequency-domain features were directly extracted from the power spectral density estimated using Welch's method using a window of 10 s with 10% overlap. The AR coefficients and model order (herein 14) were also estimated for each segment by the Berg method as described in [36]. To characterize non-stationary properties of EEG signals, thirty wavelet features were extracted from each segment using the Daubechies wavelet of order 8 (db8). We further computed six cepstral features to capture the periodicity or repeated patterns of EEG signals in neonates.

Neonatal brain signals significantly differ from adults in terms of non-stationarity, nonlinearity, and asymmetry. For example, there are alternate/discrete trace and burst patterns in the QS sleep stage. In order to identify these patterns in neonates' sleep, we used 32 features specifically adapted in the literature for sleep study in neonates [20, 37, 38]:

- The range EEG or *rEEG* is defined as the difference between maximum and minimum of the EEG segment using two-second windows with no overlap. In each 30 s epoch, 15 rEEG values were obtained and features such as mean, median, standard deviation, skewness, kurtosis, 95% and 5% percentiles, and the interquartile range were then computed.
- The Line length (LL) is a simple version of the fractal dimension, an efficient feature in detecting transient activities like burst patterns [37]. The line length is defined as the sum of the absolute values of the differences between successive instances in a predefined window. In this study, we used a one-second sliding window with a slide step of 0.12 s. The features such as mean, median, standard deviation, skewness, kurtosis, 95% and 5% percentiles, and the interquartile range were then computed for each segment.
- The line length's histogram is very efficient in identifying both QS and AS stages. The range of LL values is wide in the QS stage while the magnitude of LL values is generally significant in the AS stage. We obtained the histogram of LL values using 10 bins and extracted 17 features from the line length's histogram, such as mean, median, standard deviation, skewness, kurtosis, 95% and 5% percentiles, and the interquartile range [39].
- Four other non-linear features, Petrosian fractal dimension, mean Teager energy, mean energy, and mean curve length, were extracted from each segment [13, 35].

To reduce the risk of overfitting caused by redundancy between features, we used an efficient multivariate filter-based feature selection approach called MGCACO [26] for dimensionality reduction. In this method, a community detection algorithm is used to cluster features represented as a graph with nodes (features) and edges (similarities between nodes). The algorithm (figure S1 (available online at stacks.iop.org/JNE/17/036031/mmedia)) then iteratively selects a feature subset with minimum redundancy between features and maximum relevance to classes [26].

2.2.3. Channel selection

We further used a channel selection method to improve the classification performance and reduce the computational complexity by identifying brain areas more relevant to sleep staging in neonates. Various wrapper-based and filter-based channel selection methods have been used in EEG based applications [6, 40, 41]. We used the sequential forward selection (SFS) method, commonly used for feature selection, to select the optimal channel set using the LOOCV

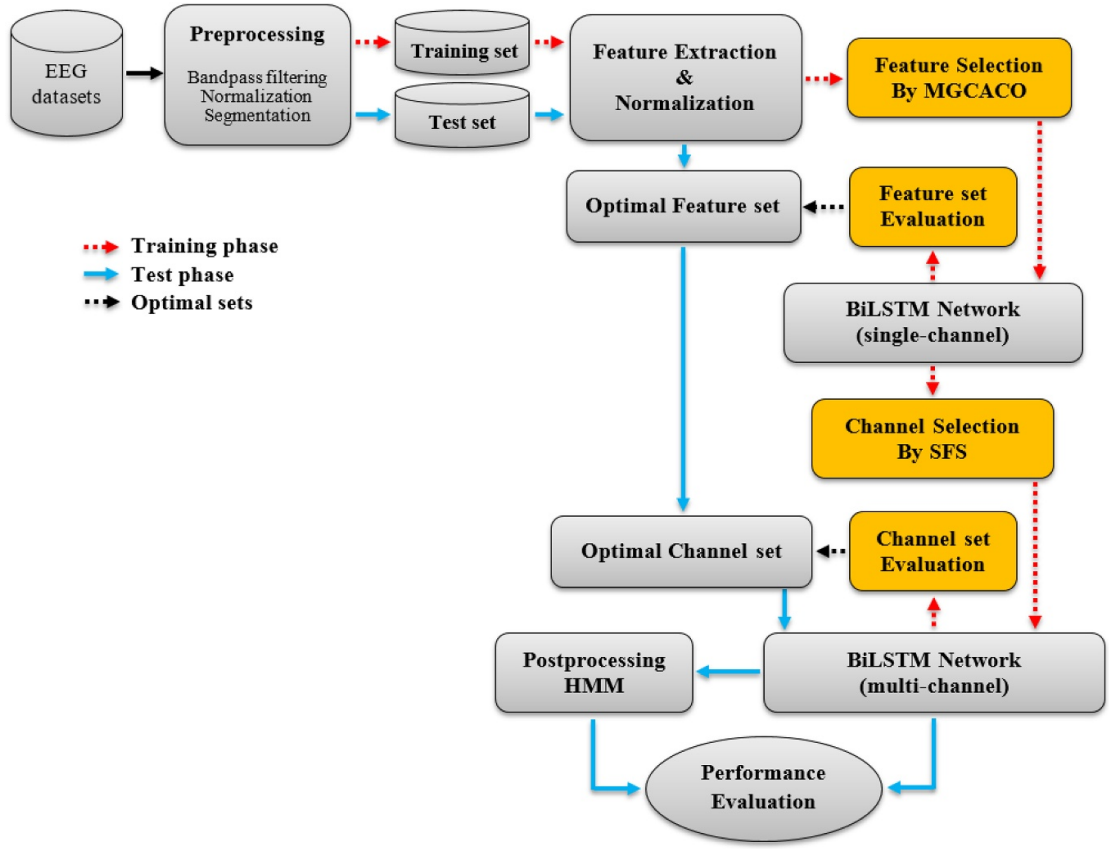

Figure 2. Block diagram of the automated algorithm designed for EEG sleep stage scoring in neonates. The data flow of the training and testing processes is shown by the red and blue arrows, respectively. In the training phase, the feature and channel selections were performed using the leave-one-out cross validation (LOOCV) strategy. The classifier was then optimized and trained using training sets. In the testing phase, the optimal feature and channel subsets were used to evaluate the classification performance on test sets using the trained classifier and hidden Markov models.

strategy as described in section 2.3.1. In this method, each of the EEG channels (16 bipolar channels or 9 mono-polar channels) was first evaluated individually, and the channel with the highest classification performance was selected as the best channel. Then, the remaining channels were added to the selected best channel one by one. The resulting channel subsets were then evaluated and the channel subset with the highest classification was selected as the optimal channel subset. This process was continued until the last channel was selected. Finally, the classification performance was plotted as a function of the number of channels. The optimal channel subsets were selected based on the maximum accuracy and the highest stability in the classification performance achieved using the minimum number of channels.

2.2.4. Classification

In the present study, we used the BiLSTM network, a powerful Recurrent Neural Network (RNN) to capture long-range dependencies between multichannel EEG signals (see supporting documents). The BiLSTM network processes the input sequence in both directions with two different hidden layers (figure S2) using the information from the past and future time steps. The generalization capability of this network can be enforced by tuning its mini-batch size [42].

In our algorithm, we defined a spatial series instead of a time series as the input of the BiLSTM network by using EEG signals from the channels from the optimal channel subset. In this manner, we assumed that at any time sample, the recorded EEG signals by the electrodes placed on the baby's head show spatial interdependence that could be modeled as a spatial series. Therefore, each 30-second epoch was fed into the BiLSTM network as a $N_f \times N_c$ feature matrix where N_f and N_c were the number of selected features and the number of selected channels, respectively. In this matrix, each epoch is represented by the optimal feature subset and the optimal channel subset arranged in rows and columns, respectively.

2.2.5. Postprocessing by HMM

We used an HMM to model the sequential nature of the sleep stages (AS_I, LVI, HVS and TA) labeled by the classifier. The aim of the HMM-based postprocessing was to incorporate the temporal structure of sleep cycles into the sleep staging procedure to reduce false positives resulting from the classification stage [12]. In this stage, the hidden states (herein, sleep

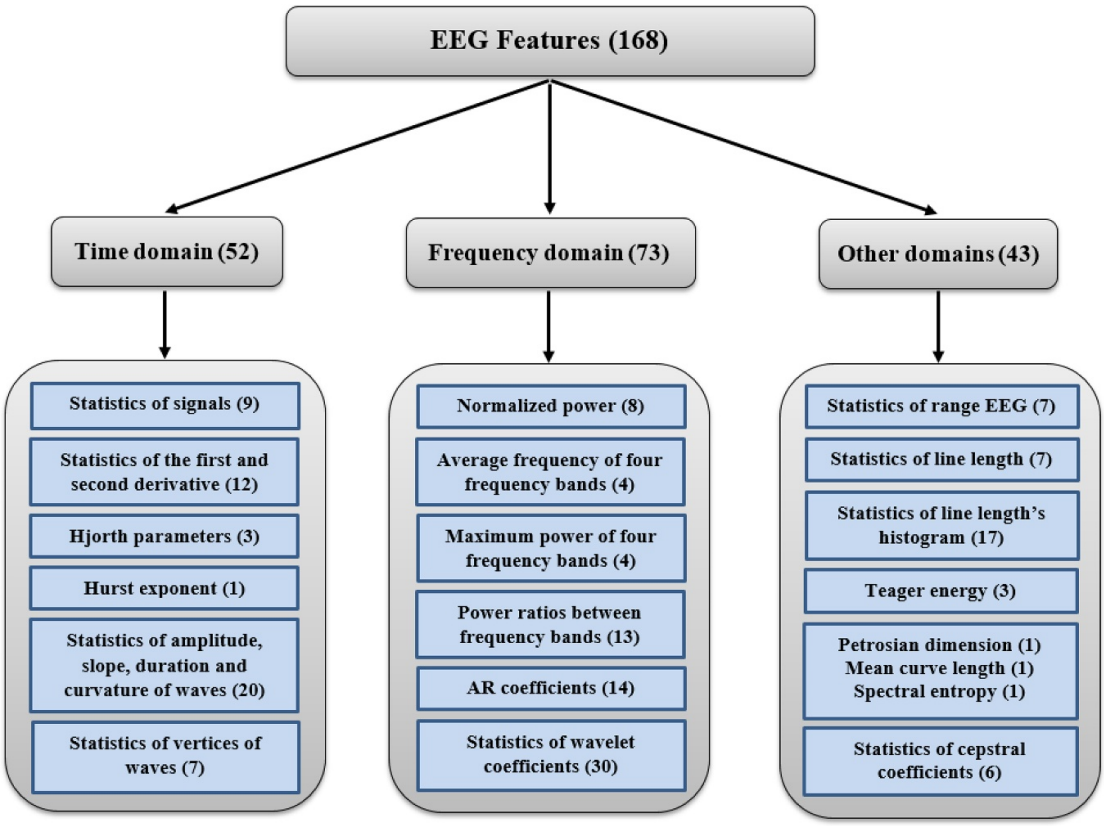

Figure 3. EEG features extracted from each EEG segment. The number of features is specified for each category.

stages) and their transitions were modeled by a first-order Markov process using the following parameters (figure 4):

- Set of states ($S_{1\times4}$ or $S_{1\times2}$): four stages $S = \mathrm{AS_I}$, HVS, TA, LVI or two stages $S = \mathrm{AS}$, QS
- Transition matrix ($T_{4\times4}$ or $T_{2\times2}$): a matrix, in which each element (i,j) defines the transition probability from the ith state to the jth state.
- Emission matrix ($E_{m\times4}$ or $E_{m\times2}$): a matrix, in which each element (i,j) defines the probability that the ith observation stays in the jth state.

2.3. Optimization procedures

2.3.1. Feature selection

We determined the optimal number of features based on the performance improvement curve obtained using the LOOCV strategy, in which the data of one subject was left out and the data of the remaining $(N_s - 1)$ subjects were used as the training set to sort the features by the MGCACO algorithm. The parameters of the MGCACO algorithm were set according to the proposed values in [26]. We determined the optimal feature subset for each channel and then a score was calculated for every single feature as follows:

$$Score(F_i) = \sum_{j=1}^{N_c} \frac{1}{s_i(j)} \quad (9)$$

Where N_c was the number of channels and $s_i(j)$ was the rank of the ith feature (F_i) on the jth channel. Then, the BiLSTM-based classifier was trained using the sorted features based on equation (9) with an increasing number of features. The trained classifier was then used to predict the class score for each segment of the training set. The receiver operating characteristic curve (ROC) curve was calculated based on the predicted class scores of the segments for each class. The area under the ROC curves (AUC) was calculated for each class and averaged over classes. Finally, the AUC curve also called the 'performance improvement curve' was computed as a function of the number of features on training sets. This procedure was repeated for all N_s runs and the average AUC curve was computed over all runs. On the AUC curve, the minimum number of features with which the average AUC of the classifier did not significantly change (less than 0.5%) was taken as the optimal number of features.

2.3.2. BiLSTM network optimization

To optimize the mini-batch size parameter as the main parameter affecting the performance of the BiLSTM network, we evaluated the classification performance improvement based on the AUC curves with increasing mini-batch size parameter using the LOOCV strategy.

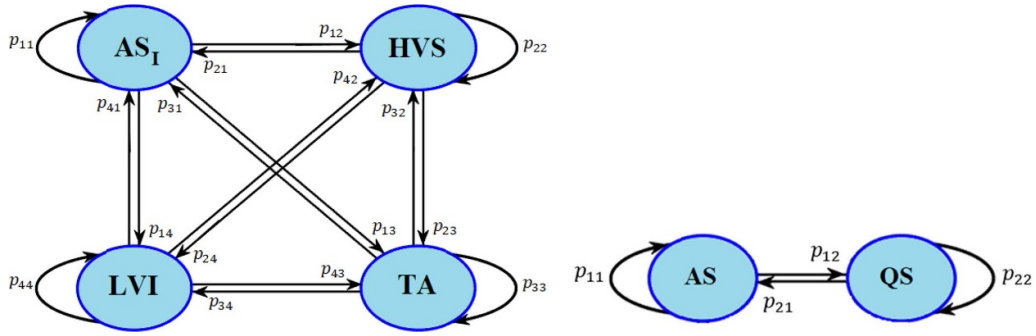

Figure 4. Fully connected HMM used for postprocessing for the four-class and two-class cases. AS_I: active sleep (sub-state 1), LVI: Low-Voltage Irregular, TA: Tracé Alternant, HVS: High Voltage Slow-wave.

2.3.3. HMM optimization

We used the LOOCV strategy to estimate the HMM parameters. At each run, one subject was left out from the dataset as the test set and the remaining subjects were used to train the first-order HMM. In our model, the predicted labels of the training set were used to train the HMM. If the total number of predicted labels was K, then the state sequence v_s and observation sequence v_o were selected as follows:

$$v_s = [L_1, L_2, \ldots, L_{K-1}]$$

$$v_o = [L_2, L_3, \ldots, L_{K-1}]$$

Where L_i is the ith label. The matrices T and E were estimated based on the vectors v_s and v_o on the training set and then the labels predicted by the BiLSTM network were refined based on the optimal HMM.

2.4. Performance evaluation

We evaluated the performance of our tool in terms of average accuracy, specificity, sensitivity, Cohen's kappa coefficient and overall accuracy for four-class (AS_I, HVS, TA, and LVI) and two-class (AS_I-LVI and HVS-TA) cases using the LOOCV and KFCV strategies mostly used by various sleeping scoring methods [2, 12]. The evaluation of the proposed method was performed in MATLAB on a PC with 8 GB RAM and a 2.7 GHz i7.

2.4.1. Leave-one-out cross-validation (LOOCV) strategy

In the LOOCV strategy, the number of runs was set to the number of subjects (N_s) . In each run, the EEG data of one neonate was selected as the test set and the data of the remaining neonates $(N_s - 1)$ were considered as the training set. Then, the evaluation parameters such as average accuracy, sensitivity, specificity, overall accuracy, and Cohen's kappa coefficient were calculated on the test data and the mean and standard deviation of the evaluation parameters were reported over all runs.

2.4.2. K-fold cross-validation (KFCV) strategy

For the KFCV strategy, we set the number of folds K to 8 in accordance with the previous study [2]. At first, the dataset was randomly divided into eight groups of two subjects. In each fold, the data of one group was selected as the test set and the remaining groups were used for training. Then, the evaluation parameters were calculated on the test set. This process was repeated five times and the algorithm was evaluated on 40 iterations (5 times $\times$ 8 folds). Finally, the mean and standard deviation of the evaluation parameters were reported over all iterations.

3. Results

3.1. Feature selection

Figure 5 shows the AUC curve as a function of the number of features for the four-class case using the monopolar and bipolar montages on the training set. To obtain these curves, all the features were sorted using the MGCACO algorithm and the LOOCV strategy. As shown, the BiLSTM classifier showed no significant performance improvement when at least 60 features were included in the feature subsets. Table 2 lists the selected features that were used for further performance evaluation.

From the top-60 most effective and relevant features (table 2) ranked by the MGCACO algorithm, thirty-three features were from the time and frequency domains. Interestingly, most of the wavelet and nonlinear features showed significant involvement in performance improvement, reflecting nonlinear and nonstationary characteristics of neonatal EEG signals. A cepstral feature was also selected, showing the rhythmicity of neonatal EEG signals.

3.2. HMM optimization

Figure 6 shows the average optimized HMMs obtained for the four-state and two-state cases. The transition and emission probabilities of the models were estimated using the LOOCV strategy. As shown, the transition probabilities between (HVS and TA) and (AS_I and HVS) are relatively stronger. In the four-state model, the infrequent sleep stage

Table 2. Most discriminant and least redundant features selected by the MGCACO algorithm, with highest occurrence across all runs.

Domain	Features	Number
Time domain	Mean, skewness and kurtosis of raw amplitudes Mean of the first and second derivative of raw amplitudes Skewness and zero-crossing of the second derivative of raw amplitudes Hjorth parameters: Complexity Median Coefficient of variation of amplitude of waves Mean, Coefficient of variation of sharpness of waves Mean of rise time/mean of fall time Mean of vertex-to-vertex amplitudes and slopes	15
Frequency domain	Normalized powers of δ_1, δ_2, α_2, β_1, β_2 Average frequency of β Maximum power of θ and α Relative Power, $\frac{\delta_1}{\delta_2}$, $\frac{\theta_1}{\theta_2}$, $\frac{\beta_1}{\beta_2}$, and $\frac{\delta}{\alpha}$, $\frac{\delta}{\beta}$, $\frac{\theta}{\alpha}$, $\frac{\theta}{\beta}$, $\frac{\alpha}{\beta}$, $\frac{\beta_1}{\alpha}$, $\frac{\beta_2}{\alpha}$	18
Wavelet domain	Skewness and zero-crossing of approximate coefficients Variance, skewness, zero-crossings and coefficient variation of the second, third and fourth level coefficients Kurtosis of the second level coefficients	15
Cepstral domain	Kurtosis of cepstral coefficients	1
Nonlinear domain	Hurst exponent Mean, median, upper margin and interquartile of range EEG Standard deviation of line length 3rd,4th and 10th bin of line length's histogram Mean and kurtosis of line length's histogram Upper margin of line length's histogram	11

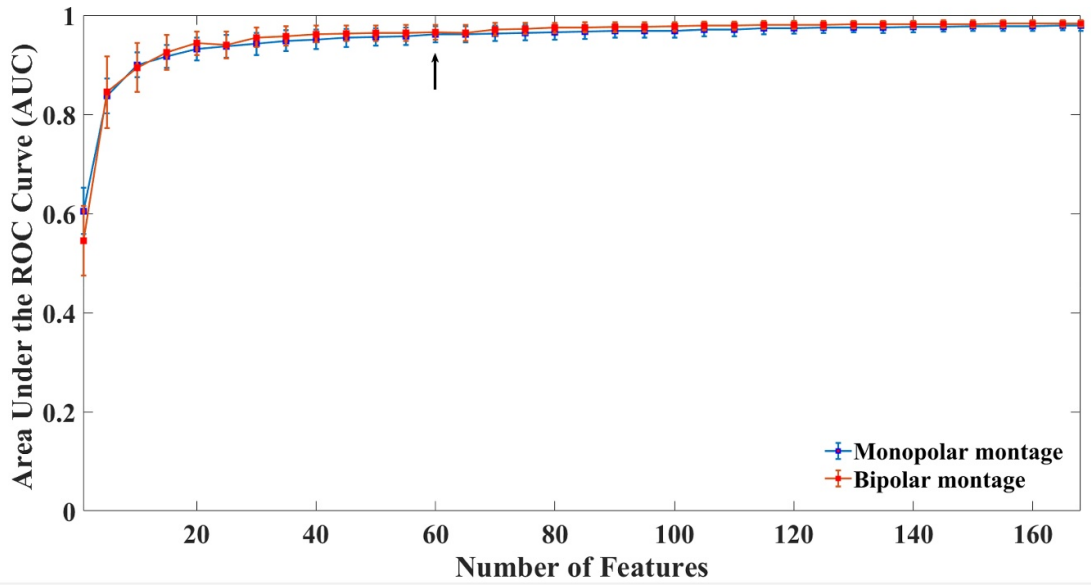

Figure 5. Average AUC curves as a function of the number of features. The AUC curves were computed using the monopolar and bipolar channels for the four-class case. The arrow indicates the optimal number of features selected for training and testing the classifier. Bars represent standard deviation.

transitions with probabilities weaker than 0.01 (1%) were eliminated during the optimization process. We used these models to reduce false positives of the sleep stage classification.

3.3. BiLSTM network optimization

To optimize the mini batch size parameter representing the number of training sets fed into the BiLSTM network at each iteration, we used the optimization procedure described in section (2.3.2). Figure 7 shows the AUC curves obtained for the BiLSTM classifier using the LOOCV strategy for the monopolar and bipolar montages. As shown, the classifier shows overfitting tendencies for the mini-batch sizes of less than 100. The classifier showed reduced performance on test sets. Moreover, the classifier performance significantly deteriorated with the min-batch size less than 200 or more than 300. We selected a mini-batch size of 200 for further performance evaluation.

3.4. Channel selection

Figure 8 shows the average AUC curves as a function of the number of channels. To sort the monopolar and bipolar channels using the training sets, the SFS

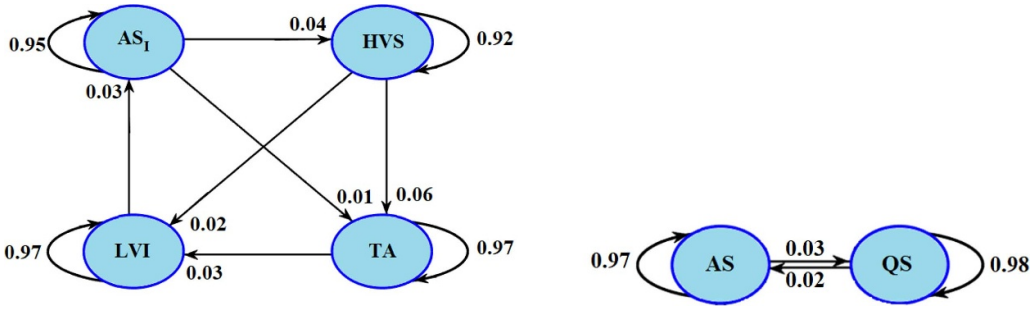

Figure 6. Hidden Markov models optimized for the four-state and two-state cases using the LOOCV strategy. In the four-state model, inter-state transitions with probabilities weaker than 0.01 (1%) were considered as infrequent transitions and eliminated. AS_I: active sleep (sub-state 1), LVI: Low-Voltage Irregular, TA: Tracé Alternant, HVS: High Voltage Slow-wave.

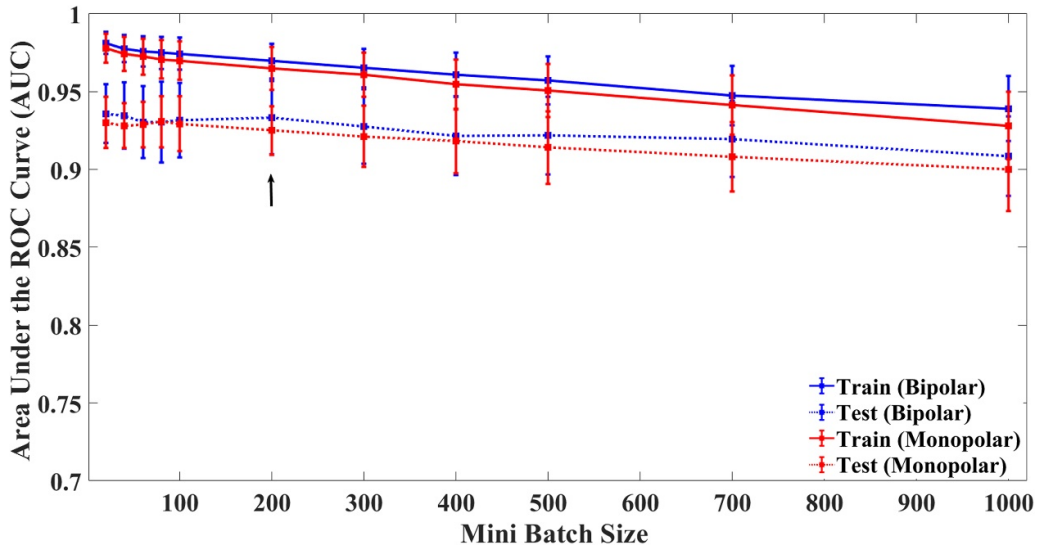

Figure 7. Average AUC curves as a function of the mini batch size. The AUC curves were computed using the monopolar and bipolar montage on training (solid line) and test sets (dashed line) for the four-class case. The arrow indicates the optimal batch size selected to increase the generalization capability of the classifier. Bars represent standard deviation.

algorithm and the LOOCV strategy were used. Table 3 lists the channels sorted based on their performance. As shown, the classifier reached the highest stability in performance and maximum accuracy when at least six mono or bipolar channels were selected for sleep stage scoring. The optimal channels were mostly selected from both brain hemispheres, indicating asymmetry in EEG characteristics in neonates. The channel selection results clearly emphasize the importance of multichannel EEG sleep staging in neonates.

3.5. Performance evaluation

3.5.1. LOOCV strategy

Tables 4 and 5 report the per-channel average accuracy, sensitivity, specificity, Cohen's kappa coefficient, and overall accuracy for the four-class case using the monopolar and bipolar channels, respectively. As shown, the classification results obtained using Fp1 and Fp2 and Fp1-T3 and Fp2-T4 were more significant than those achieved using other channels. This suggests that frontal regions provide more discriminant information on different sleep stages in neonates. Moreover, the sleep staging showed a significant improvement using bipolar channels in comparison with the referential montage.

In tables 6 and 7, the multi-channel average accuracy, sensitivity, specificity, Cohen's kappa coefficient, and overall accuracy on all classes are reported for the four-class case based on the LOOCV strategy. To obtain these results, we used the first six most significant monopolar and bipolar channels based on the results reported in section 3.4. As shown, the overall accuracy of the proposed algorithm improved up to 4.5% using the multichannel approach in comparison with the single-channel evaluation. This suggests that bilateral electrodes carry different relevant information required to improve the sleep staging performance in neonates regardless of the number of sleep stages. The HMM-based postprocessing improved the performance using both montages up to 3% and 1.5% for the four-class and two-class cases, respectively.

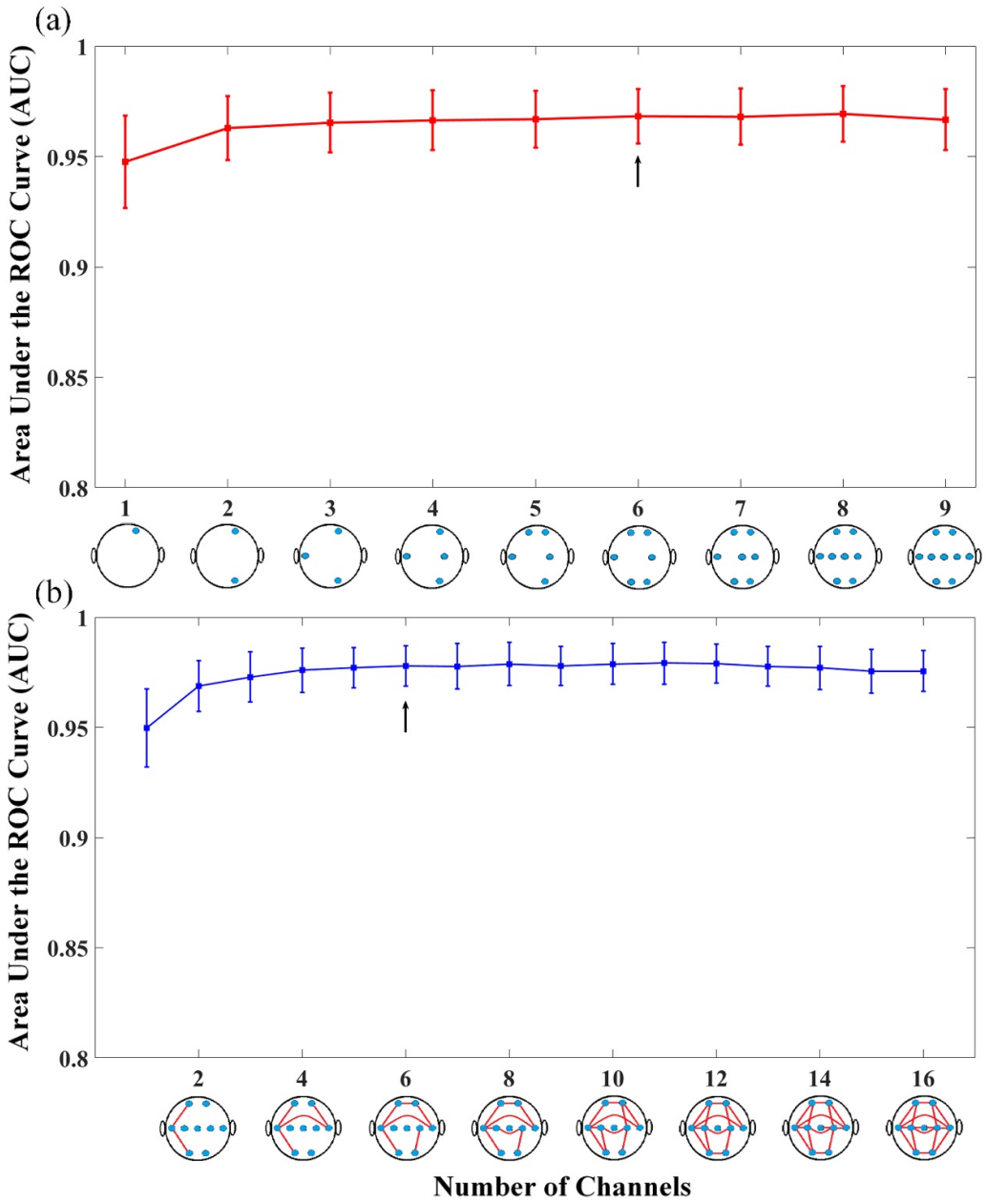

Figure 8. AUC curves as a function of the number of channels for the monopolar (a) and bipolar (b) montage for the four-class case. The arrows indicate the optimal channel subsets used to achieve the maximum accuracy and the highest stability in performance for the classifier using the minimum number of channels. Bars represent standard deviation.

3.5.2. KFCV strategy

Figure 9 shows confusion matrices computed using the KFCV strategy at the classification and the HMM-based postprocessing levels for the four-class case using the most significant monopolar and bipolar channels. As shown, the postprocessing stage could improve the overall accuracy by 3% for both montages. The HMM increased the average sensitivity (figures 9(a) and (b)) by 6% for all sleep stages except for AS_I using the monopolar montage. The postprocessing could also reduce false positives (figures 9(c) and (d)) efficiently for all stages using the bipolar montage even for AS_I. The maximum improvement in sensitivity was obtained for TA for both montages. Misclassification mostly occurred between pairs AS_I-HVS and AS_I -TA. The HMM-based postprocessing helped to significantly decrease false positives for AS_I (monopolar) and TA (bipolar).

Tables 8 and 9 show the multi-class average accuracy, sensitivity, specificity, Cohen's kappa coefficient, and overall accuracy of the proposed method for the four-class case based on the KFCV strategy. The algorithm's performance based on the KFCV strategy slightly improved in comparison with the LOOCV strategy.

4. Discussion

In this paper, we developed an automated multichannel approach based on the BiLSTM network

Table 3. Monopolar and bipolar channels ranked by the SFS algorithm.

Rank	Mono-polar	AUC	Bipolar	AUC
1	Fp2	0.9476	Fp1-C3	0.9496
2	O2	0.9629	T3-O1	0.9687
3	T3	0.9654	C4-T4	0.9729
4	C4	0.9664	O1-O2	0.9760
5	Fp1	0.9670	Fp2-T4	0.9771
6	O1	0.9683	T3-T4	0.9783
7	Cz	0.9680	Fp1-Fp2	0.9777
8	C3	0.9693	Fp1-T3	0.9787
9	T4	0.9667	T3-C3	09779
10	–	–	Fp2-C4	0.9785
11	–	–	C3-C4	0.9791
12	–	–	T3-O1	0.9787
13	–	–	C3-O1	0.9767
14	–	–	C3-O1	0.9770
15	–	–	Cz-O2	0.9756
16	–	–	T4-O2	0.9756

for EEG sleep staging in neonates. To improve the classification performance, the method was optimized in three steps. First, a subset of linear and nonlinear features was selected based on the redundancy analysis and relevancy analysis by using the MGCACO algorithm [26]. Then, an optimal sequence of bipolar/monopolar channels was selected by the SFS method. Finally, a first-order HMM was used for postprocessing to reduce false detections by modelling the transitions between sleep stages.

So far, a few neonatal sleep staging methods have been developed with varying degree of success [2, 18–21, 23, 43]. Using a combination of features extracted from EEG, electromyography (EMG), electro-oculography (EOG), respiratory (PNG) and electrocardiography (ECG) signals, Gerla *et al* have developed two automatic sleep stage classification approaches [44, 45]. In their first approach, they have developed a multichannel method based on the HMM combined with expectation maximization (EM) algorithm, reaching an average classification accuracy of 82% on three classes, Wake, Active and Quiet sleep in term neonates. The main advantage of this approach is the characterization of the sleep and wake stages using five polysomnography signals required to differentiate wakefulness from other sleep stages. However, they have only used one frequency-domain feature, which limits the efficiency of the method for sleep stage classification using neonatal EEG signals exhibiting nonstationary and nonlinear characteristics [46]. In their later work, they obtained an accuracy of 68.8% and 66.2% using the ACO-Dtree and random tree classifiers, respectively, using power spectral features extracted from polysomnographic recordings to differentiate four neonatal behavioral states: quiet sleep, active sleep, wakefulness and movement artifact [45].

Using a set of features and Fisher's linear discriminant classifier, Löfhede *et al* developed an automatic classification of background EEG activity in full-term healthy newborn babies [47]. They reported an accuracy of 100% when separating burst suppression EEG from other stages (active/quiet awake, active/quiet sleep) and a true positive rate of 93% when separating quiet sleep from the other types. Fraiwan *et al* presented an automated sleep staging method based on entropy features extracted from time-frequency distributions of EEG recordings from full-term neonates [18]. They have achieved a classification accuracy of 84% and a kappa coefficient of 0.65 using Wigner–Ville distribution-based features and artificial neural networks. Using time-frequency features and Support Vector Machines (SVM), Čić *et al* presented an approach for single-channel sleep stage classification of daytime sleep in neonates with an average accuracy of 80% [21]. Fraiwan and Lweesy presented a single-channel classification method for automatic sleep stage scoring in neonates based on multiscale entropy with an accuracy of 81.3% and a kappa coefficient of 66.7% to classify three states (awake, active, quiet) of newborns sleep at term age [19]. De Wel *et al* have developed a multichannel and multiscale entropy approach for QS-non-QS classification with a sensitivity and specificity of 90% [25].

In a more comprehensive study, Koolen *et al* [20] proposed a multichannel algorithm for sleep stage classification of active and quiet sleep in term babies using a set of 57 EEG features extracted in spatial, temporal, and spectral domains. They have used a greedy algorithm for feature selection to reduce the feature space dimensionality. They have reported an accuracy of 85% to classify quiet and active sleep epochs. A cluster-based adaptive sleep staging (CLASS) method has been developed for QS detection in preterm and term babies using adaptive segmentation and nine frequency and time-domain features [43]. The CLASS method achieved a mean kappa of 0.62 ($\pm$0.19) in term neonates.

For performance evaluation, the majority of the aforementioned methods have used the KFCV strategy [2, 19–21, 23, 25, 44], in which the dataset is randomly divided into k subsets (folds) of equal size. At each of k runs, one of the k subsets is used for testing the classifier and the remaining k-*1* subsets are used for training. Finally, the average performance parameters are computed over all k runs. In [19, 21, 25, 44], k has been set to 10 to generate training and test sets using EEG segments pooled together from all neonates included for sleep stage classification. This strategy has been mainly used to ensure the best possible accuracy for sleep stage classification in neonates in an intra-subject manner. However, the generalization ability of these methods might be very limited across different datasets.

In our study, we used an 8-fold cross-subject validation strategy [2, 23], which is effective in better illustrating how well the method performs when making predictions on unseen sleep data. In [2],

Table 4. Average per-channel Cohen's kappa Coefficient, accuracy, sensitivity, specificity and overall accuracy of the proposed method at the classification and postprocessing level. The evaluation results were obtained using the LOOCV strategy and monopolar channels for the four-class case.

Test set	Postprocessing step	Kappa	Accuracy (%)	Sensitivity (%)	Specificity (%)	Overall Accuracy (%)
Fp1	Off	0.64	86.93	71.36	91.16	73.86
	On	**0.70**	**89.17**	**75.65**	**92.64**	**78.33**
Fp2	Off	0.64	86.81	71.08	91.06	73.62
	On	0.69	88.76	74.89	92.38	77.52
C3	Off	0.56	83.92	65.21	89.09	67.84
	On	0.61	85.85	68.95	90.39	71.70
C4	Off	0.55	83.54	64.73	88.83	67.07
	On	0.59	85.22	67.90	89.99	70.44
Cz	Off	−0.01	64.59	24.21	74.61	29.18
	On	−0.02	64.95	23.78	74.41	29.90
T3	Off	0.59	85.20	67.68	89.97	70.39
	On	0.66	87.51	72.05	91.50	75.02
T4	Off	0.59	85.16	67.83	89.95	70.32
	On	0.66	87.48	72.35	91.51	74.96
O1	Off	0.59	84.86	67.54	89.72	69.71
	On	0.65	87.27	72.21	91.34	74.54
O2	Off	0.60	85.26	68.65	89.98	70.51
	On	0.66	87.56	73.09	91.55	75.13
Average on channels	Off	**0.53**	**82.92**	**63.14**	**88.26**	**65.83**
	On	**0.58**	**84.86**	**66.76**	**89.52**	**69.73**

Table 5. Average per-channel Cohen's kappa Coefficient, accuracy, sensitivity, specificity and overall accuracy of the proposed method at the classification and postprocessing level. The evaluation results were obtained using the LOOCV strategy the bipolar channels for the four-class case.

Test set	Postprocessing step	Kappa	Accuracy (%)	Sensitivity (%)	Specificity (%)	Overall Accuracy (%)
Fp1-Fp2	Off	0.61	85.73	68.51	90.42	71.47
	On	0.66	87.75	72.48	91.75	75.51
T3-C3	Off	0.61	85.63	68.38	90.27	71.27
	On	0.68	88.24	73.29	92.00	76.47
C3-Cz	Off	0.55	83.67	64.88	88.92	67.33
	On	0.60	85.45	68.46	90.13	70.90
Cz-C4	Off	0.55	83.51	64.89	88.82	67.01
	On	0.59	85.22	68.21	89.98	70.44
C3-C4	Off	0.59	84.97	67.76	89.85	69.93
	On	0.64	86.69	70.96	91.03	73.39
C4-T4	Off	0.57	84.40	66.41	89.41	68.80
	On	0.62	86.32	70.25	90.72	72.63
T3-T4	Off	0.60	85.43	68.08	90.13	70.86
	On	0.67	87.84	72.79	91.77	75.67
O1-O2	Off	0.53	82.85	63.19	88.35	65.70
	On	0.60	85.27	67.83	90.03	70.53
Fp1-C3	Off	0.65	87.15	71.72	91.32	74.29
	On	0.70	89.21	75.92	92.72	78.42
C3-O1	Off	0.59	84.88	67.67	89.77	69.77
	On	0.65	87.12	72.23	91.29	74.23
Fp2-C4	Off	0.66	87.51	72.78	91.54	75.01
	On	0.71	89.37	76.44	92.80	78.73
C4-O2	Off	0.60	85.34	68.55	90.03	70.68
	On	0.66	87.54	72.76	91.54	75.08
Fp1-T3	Off	0.65	87.16	71.32	91.34	74.32
	On	**0.72**	**89.63**	**76.26**	**93.02**	**79.26**
T3-O1	Off	0.58	84.84	66.90	89.70	69.67
	On	0.66	87.68	72.38	91.63	75.37
Fp2-T4	Off	0.65	87.08	71.37	91.28	74.16
	On	0.71	89.58	76.21	92.93	79.16
T4-O2	Off	0.56	84.00	65.62	89.16	68.00
	On	0.63	86.30	70.20	90.75	72.61
Average on channels	Off	**0.60**	**85.26**	**68.00**	**90.02**	**70.52**
	On	**0.66**	**87.45**	**72.29**	**91.50**	**74.90**

Table 6. Average multi-channel Cohen's kappa Coefficient, accuracy, sensitivity, specificity and overall accuracy of the proposed method based on the LOOCV strategy at the classification and postprocessing level for the four-class case.

Montage	Postprocessing stage	Kappa	Accuracy (%)	Sensitivity (%)	Specificity (%)	Overall Accuracy (%)
Mono-polar	Off	0.6745	88.14	73.72	91.94	76.28
	On	0.7152	89.64	76.50	92.94	79.29
Bipolar	Off	0.7188	77.57	77.57	93.08	79.43
	On	**0.7585**	**80.37**	**80.37**	**94.06**	**82.35**

Table 7. Average multi-channel Cohen's kappa Coefficient, accuracy, sensitivity, specificity and overall accuracy of the proposed method based on the LOOCV strategy at the classification and postprocessing level for the two-class case.

Montage	Postprocessing stage	Kappa	Accuracy (%)	Sensitivity (%)	Specificity (%)	Overall Accuracy (%)
Mono-polar	Off	0.8513	92.59	94.83	90.17	92.59
	On	0.8780	93.92	96.80	**90.81**	93.92
Bipolar	Off	0.8676	93.40	96.19	90.39	93.40
	On	**0.8852**	**94.29**	**98.21**	90.05	**94.29**

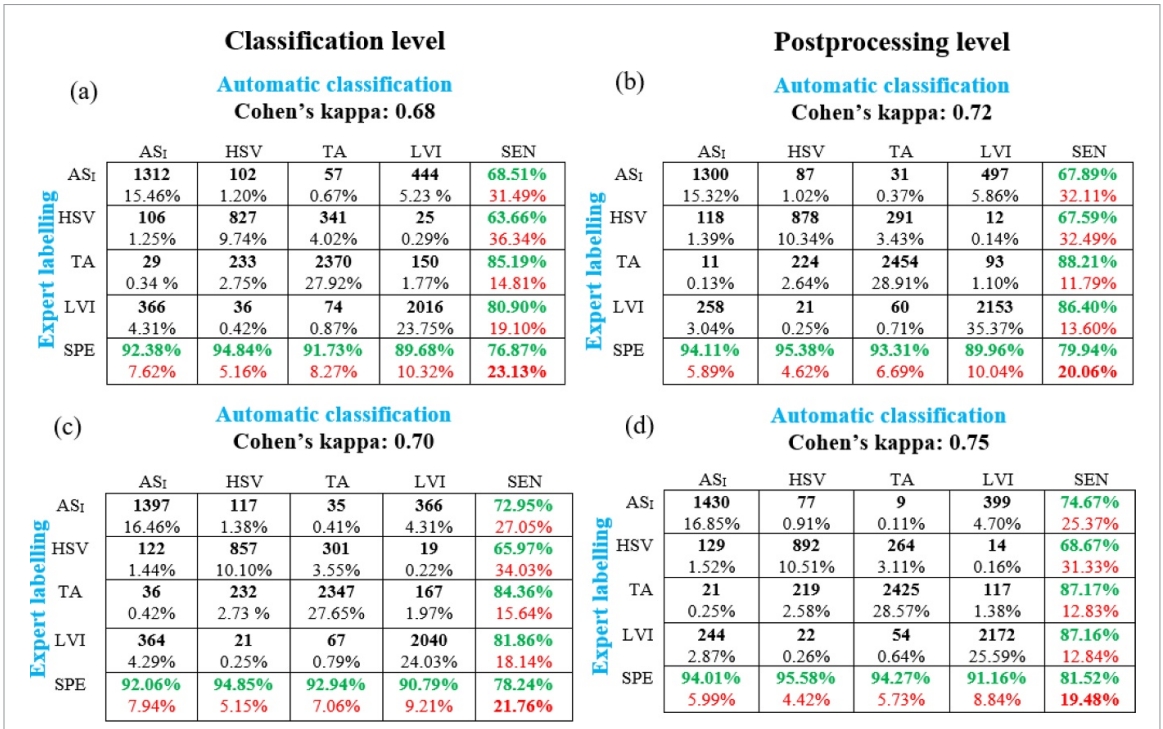

Figure 9. Confusion matrix obtained by the multichannel method using the monopolar (a) and (b) and bipolar (c) and (d) montages. The evaluation results were obtained using the KFCV strategy for the four-class case. Each column represents the classification results for the corresponding class. Each row represents the actual class. The number and percentage of correct classifications are presented in the diagonal cells, while the other cells represent the misclassified predictions. The last column represents the recall (sensitivity) rate of the classifier for each class. The last row represents the specificity of the classifier for predicting each class. The last diagonal cell represents the overall accuracy and false positive rate. SEN: sensitivity, SPE: specificity.

Table 8. Average multichannel Cohen's kappa Coefficient, accuracy, sensitivity, specificity and overall accuracy of the proposed method based on the KFCV strategy at the classification and postprocessing level for the four-class case.

Montage	Postprocessing step	Kappa	Accuracy (%)	Sensitivity (%)	Specificity (%)	Overall Accuracy (%)
Monopolar	Off	0.66 ± 0.09	87.88 ± 3.90	73.76 ± 7.90	91.87 ± 2.32	75.76 ± 6.59
	On	0.71 ± 0.09	89.44 ± 3.16	76.56 ± 7.96	92.88 ± 2.27	78.88 ± 6.33
Bipolar	Off	0.70 ± 0.10	89.28 ± 3.49	76.89 ± 7.70	92.84 ± 2.41	78.55 ± 6.98
	On	$\mathbf{0.75 \pm 0.09}$	$\mathbf{90.84 \pm 3.36}$	$\mathbf{76.66 \pm 7.93}$	$\mathbf{93.84 \pm 2.36}$	$\mathbf{81.68 \pm 6.72}$

Pillay *et al* presented a method based on the Gaussian mixture model (GMM) and HMM and the Minimum Redundancy Maximum Relevance (mRMR) feature selection to differentiate four AS and QS sub-states using the same sleep EEG data used in our study. They have reported a mean kappa (±standard deviation) of 0.62 (±0.16) for the HMM compared to the GMM (0.55 ± 0.15) using the monopolar montage. In a very

Table 9. Average multichannel Cohen's kappa Coefficient, accuracy, sensitivity, specificity and overall accuracy of the proposed method based on the KFCV strategy at the classification and postprocessing level for the two-class case.

Montage	Postprocessing step	Kappa	Accuracy (%)	Sensitivity (%)	Specificity (%)	Overall Accuracy (%)
Monopolar	Off	0.86 ± 0.12	93.00 ± 5.82	94.41 ± 3.51	91.85 ± 9.77	93.00 ± 5.82
	On	0.89 ± 0.11	94.55 ± 5.47	96.62 ± 2.22	$\mathbf{92.59 \pm 9.66}$	94.55 ± 5.47
Bipolar	Off	0.88 ± 0.12	93.90 ± 5.84	95.76 ± 2.25	92.28 ± 11.49	93.90 ± 5.84
	On	$\mathbf{0.90 \pm 0.13}$	$\mathbf{95.10 \pm 6.56}$	$\mathbf{98.02 \pm 1.18}$	92.09 ± 12.89	$\mathbf{95.10 \pm 6.56}$

recent study, Ansari *et al* [23] have used a convolutional neural network (CNN) for sleep stage classification in preterm and term neonates. The network has been directly trained with down-sampled 30 s EEG segments from the same neonatal EEG dataset used in our study. Their method has achieved a mean kappa coefficient, sensitivity, and specificity of 0.66, 72%, and 92% for the four-class case, respectively [23]. Using the same cross-subject validation strategy [2, 23], we evaluated the performance of the BiLSTM network trained with the input feature matrices representing multichannel EEG signals. We achieved a mean kappa and an overall accuracy of 0.71 and 78.9% for the four-class case using the monopolar montage. Our method showed an average improvement of 4% and 3% in overall accuracy in comparison with the accuracies reported in [2] and [23], respectively. Using the bipolar montage, we further increased the mean kappa and the overall accuracy of our method to 0.75 and 81.7%, respectively, i.e. 6% and 5% improvement in overall accuracy compared to the overal accuraccies of the methods presented by Pillay *et al* [2] and Ansari *et al* [23].

The LOOCV strategy has also been used in a few studies [20, 21, 44] to evaluate the generalization ability of the sleep stage classification methods. The LOOCV strategy is k-fold cross-validation with k equal to N_s, the total number of subjects in the dataset. Using this strategy, our approach obtained a mean kappa and an overall accuracy of 0.72 (0.76) and 79.3% (82.4%), respectively, using the monopolar (bipolar) montage. The kappa coefficients of our method are reasonable as the manual scoring accuracy remains between 70% and 80% even between neurologists [45]. In comparison with the multi-channel approaches evaluated using the LOOCV strategy [20, 44], our results achieved significant improvements in accuracy up to 9% and 12% for the two-class and four-class cases, respectively. However, it is difficult to directly compare our evaluation results with those reported in these approaches due to the differences in datasets, number of neonates, number of sleep stages and EEG montages used for sleep stage classification.

We achieved an average accuracy of 95.10% for the scoring of active and quiet sleep stages. The classification accuracy increases in the two-state case when each state is characterized by specific patterns of changes, for example, burst suppression is very distinct from TA, which is characterized by more periodic EEG activity interrupted with low amplitude inter-burst activities [48]. The classification performance deteriorates using only EEG signals when the overlap in frequency and amplitude between states like AS_{II} and Wake [2]. This is why we did not consider a separate state for Wake showing EEG characteristics similar to those observed in AS_{II} [2]. To distinguish these two states, other information such as eyes open/closed, and changes in body movement is required [3, 32].

The most relevant and least redundant features selected by the MGCACO algorithm were mostly from those extracted in the time, frequency, and nonlinear domains. Compared to the features selected for adult sleep stage scoring in our previous work [12], nonlinear and wavelet features showed more discrimination power in the present study, indicating non-stationary and nonlinear nature of EEG signals in neonates.

Interestingly, compared to forty features selected in our previous study for adult sleep stage classification, sleep scoring in neonates required twenty more features (up to sixty features in total) to improve the performance and reach stability in accuracy. Our results are in line with those reported by [2], who have shown that increasing the number of features up to 30 could improve the classification performance significantly using HMM and GMM. They have also found that the most discriminant features selected by mRMR were those derived from wavelet decompositions and empirical mode decompositions (EMDs). This finding shows the non-stationary and nonlinear nature of the sleep EEG in neonates [2].

The channel selection method used in our approach selected channels from both hemispheres, including bipolar channels Fp1-C3, T3-O1, C4-T4, O1-O2, FP2-T4, and T3-T4. This result clearly reflects the asymmetry in EEG background activity between left and right hemispheres and between frontal and occipital regions. Our channel selection results are perfectly in line with those reported by [6], who have used a multivariate analysis approach for the optimal selection of the EEG channels for sleep staging in neonates. They have also shown that among all bipolar channels, sleep staging using Fp1-T3, T3-O1, Fp2-T4, T4-O2, and Fp1-C3 achieved the highest performance in comparison with other channels. They have, however, used a time consuming exhaustive

search method to evaluate all channel combinations. We have used the SFS method, which is a relatively low-cost algorithm for channel selection. In our algorithm, the channel selection procedure reduced the computational complexity and improved the classification performance by 2%. This is in contrast with the results reported by [20], who have shown that reducing the number of EEG channels from eight to 4 channels (Fp1, Fp2, C3, and C4) could decrease the accuracy by 2.2%. Our approach also achieved more performance improvement using bipolar channels compared to the referential ones. Moreover, the overall accuracy of the proposed algorithm was improved up to 4.5% using the multichannel approach in comparison with the single-channel evaluation. Our results confirm the superiority of the multichannel approaches over the single-channel methods for EEG-based sleep staging in neonates [19–21, 25]. This finding is also in line with the guidelines provided by André *et al* [5], who have recommended the use of multichannel monopolar or bipolar EEG signals as mimimum requirement for sleep scoring in neonates.

Ansari *et al* [23] have also shown the improvement in sleep staging performance using the multichannel approach in term neonates. This finding emphasizes the importance of multichannel sleep staging for performance improvement in neonates. For older babies (aged 3 months), single-channel sleep staging has shown similar performance using different EEG channels [21].

Overall, the HMM-based postprocessing stage could improve the overall accuracy by up to 3% for all stages by reducing false positives mostly for HVS (monopolar montage) and LVI (bipolar montage). Most of the misclassification occurred between pairs HVS-TA and AS_I-LVI. Our results are perfectly consistent with the findings reported by [2] and [23]. In our study, the HMM could model frequent transitions from HVS to TA (as shown in figure 6). The infrequenct transitions (e.g. from TA to HVS) were then removed through the optimization process. Moreover, the high self-transition probabilities could significantly reduce false positives occurring within the states [2].

In general, deep neural networks do not require feature extraction/selection and can work directly with raw data [23, 49]. However, they require large training data with considerable computational resources [49], which make them computationally more expensive in comparison with other classifiers, especially for real-time applications. To reduce the computational cost of the sleep staging method presented in the present study, we used feature matrices representing multichannel EEG signals to train the BiLSTM network. By doing so, we considerably reduced the amount of training data all by keeping the accuracy rates high. We further used feature and channel selection methods to reduce the dimensionality of the input data and consequently the computational cost. Practically speaking, once trained, the BiLSTM network was relatively cheap in terms of computational cost during the test phases. The only time-consuming part is feature extraction, especially for real-time applications.

4.1. Limitations and future work

Our work has two main limitations. In the neonatal period, active and quiet sleep are largely distinguishable based on their EEG patterns [5]. The indeterminate periods, however, show non-EEG characteristics of AS and EEG characteristics of QS [2, 3]. In healthy term neonate, the wakefulness state is characterized by eyes open, continuous EEG background with low to medium amplitudes and mixed frequencies (predominantly theta and delta with overriding beta activity), irregular respirations, and spontaneous movements of the limbs and body [4]. The active sleep also exhibits an EEG background activity called 'activité moyenne', indistinguishable from that of normal wakefulness. This stage is more characterized by eyes closed, intermittent REM periods, and irregular respirations with small and large body movements [4]. Different from the other two states, quiet sleep is clinically characterized by eye closure, absent REM, and scant body movements [4]. In this state, the EEG background activity called 'tracé alternant' shows burst activities with high amplitudes (50–150 μV peak-to-peak) predominantly of delta activities that alternate with short periods of interburst mixed theta and delta activities with lower amplitudes (25–50 μV peak-to-peak) [4]. In term age, the wakefulness and REM sleep and transitions between them are particularly difficult to identify using only EEG signals due to similarities in background activities, both exhibiting LVI or continuous mixed frequencies patterns [3–5]. This is why in addition to EEG, other physiological and behavioral variables including EOG, EMG, vocalization (whimpering, crying) as well as the irregularity/regularity of respiration [3–5] are required to automatically differentiate between sleep states and wakefulness. In the present study, we focused on four states (AS_I, LVI, HVS, and TA) for sleep stage classification in full-term neonates using only multichannel EEG signals. Like other existing methods developed for automatic sleep stage classification focusing mainly on sleep-only cycles [2, 20, 23, 43], the application of our method remains limited for long-term sleep stage scoring in neonatal intensive care units, in which neurologists need to evaluate brain function over the whole sleep-wake cycles, especially during early brain development.

Second, we only included EEG data from full-term healthy neonates aged between 38 and 42 weeks of postmenstrual age (PMA) in our study. To gain a full picture of neurodevelopment with regard to the

sleep structure, further work is required to evaluate the performance of the tool using polysomnographic recordings from premature babies and infants older than 42 weeks PMA. Moreover, the level of brain maturity and the degree of symmetry in amplitude and frequency between the two hemispheres should be investigated by exploring the evolution of spectral, nonlinear, and complexity features over the course of neurodevelopment [4, 25, 46]. Connectivity features are also of great importance to explore the degree of synchrony between the brain hemispheres as it has been reported that the degree of inter-hemispheric synchrony increases from approximately 30 weeks PMA until term when both hemispheres are completely synchronous [4]. The limitations in the present work highlight an important focus for future investigations.

5. Conclusion

In this paper, we presented an automated multichannel approach based on the BiLSTM network and HMM as postprocessing for EEG sleep stage scoring in neonates. The classification performance was significantly improved by using an efficient evolutionary feature selection method developed based on the relevance and redundancy analysis. The multichannel approach was further optimized by using a channel selection method which efficiently reduced the dimensionality at the spatial scale. Finally, the HMM-based postprocessing stage reduced false positives by incorporating the knowledge of between and within stage probabilities into the classification process. The proposed method showed significant performance improvement in comparison with other sleep staging methods reported in the literature. Our future work will aim at using and optimizing the presented approach for sleep stage scoring in preterm babies.

Acknowledgments

The authors would like to thank the parents and infants involved in this study and the staff at the UZ Leuven NICU. This research was funded by the Wellcome Trust (grant 095802/B/110Z), IWT Leuven Belgium (grant TBM 110697-NeoGuard), and Bijzonder Onderzoeksfonds KU Leuven (BOF): The effect of perinatal stress on the later outcome in preterm babies (grant C24/15/036). This work was partially supported by the Cognitive Science and Technology Council (CSTC) of Iran under the Grant Nos. 1896 and 3465.

ORCID iD

Ardalan Aarabi https://orcid.org/0000-0001-5141-9248

References

[1] Silber M H, Ancoli-Israel S, Bonnet M H, Chokroverty S, Grigg-Damberger M M, Hirshkowitz M, Kapen S, Keenan S A, Kryger M H and Penzel T 2007 The visual scoring of sleep in adults *J. Clin. Sleep Med.* **3** 22

[2] Pillay K, Dereymaeker A, Jansen K, Naulaers G, Van Huffel S and De Vos M 2018 Automated EEG sleep staging in the term-age baby using a generative modelling approach *J. Neural Eng.* **15** 36004

[3] Grigg-Damberger M M 2016 The visual scoring of sleep in infants 0 to 2 months of age *J. Clin. Sleep Med.* **12** 429–45

[4] Tsuchida T N, Wusthoff C J, Shellhaas R A, Abend N S, Hahn C D, Sullivan J E, Nguyen S, Weinstein S, Scher M S and Riviello J J 2013 American clinical neurophysiology society standardized EEG terminology and categorization for the description of continuous EEG monitoring in neonates: report of the American clinical neurophysiology society critical care monitoring committee *J. Clin. Neurophysiol.* **30** 161–73

[5] André M, Lamblin M-D, d'Allest A-M, Curzi-Dascalova L, Moussalli-Salefranque F, Nguyen The Tich S, Vecchierini-Blineau M-F, Wallois F, Walls-Esquivel E and Plouin P 2010 Electroencephalography in premature and full-term infants. Developmental features and glossary *Neurophysiol. Clin. Neurophysiol.* **40** 59–124

[6] Piryatinska A, Woyczynski W A, Scher M S and Loparo K A 2012 Optimal channel selection for analysis of EEG-sleep patterns of neonates *Comput. Methods Programs Biomed.* **106** 14–26

[7] Stockard-Pope J E, Werner S S and Bickford R G 1992 *Atlas of Neonatal Electroencephalography* (New York: Raven Press)

[8] Hassan A R and Bhuiyan M I H 2017 Automated identification of sleep states from EEG signals by means of ensemble empirical mode decomposition and random under sampling boosting *Comput. Methods Programs Biomed.* **140** 201–10

[9] Hassan A R and Bhuiyan M I H 2016 A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features *J. Neurosci. Methods* **271** 107–18

[10] Seifpour S, Niknazar H, Mikaeili M and Nasrabadi A M 2018 A new automatic sleep staging system based on statistical behavior of local extrema using single channel EEG signal *Expert Syst. Appl.* **104** 277–93

[11] Sharma R, Pachori R B and Upadhyay A 2017 Automatic sleep stages classification based on iterative filtering of electroencephalogram signals *Neural Comput. Appl.* **28** 2959–78

[12] Ghimatgar H, Kazemi K, Helfroush M S and Aarabi A 2019 An automatic single-channel EEG-based sleep stage scoring method based on hidden Markov model *J. Neurosci. Methods* 108320

[13] Aboalayon K A I, Faezipour M, Almuhammadi W S and Moslehpour S 2016 Sleep stage classification using EEG signal analysis: a comprehensive survey and new investigation *Entropy* **18** 272

[14] Fonseca P, Den Teuling N, Long X and Aarts R M 2018 A comparison of probabilistic classifiers for sleep stage classification *Physiol. Meas.* **39** 55001

[15] Fonseca P, Long X, Radha M, Haakma R, Aarts R M and Rolink J 2015 Sleep stage classification with ECG and respiratory effort *Physiol. Meas.* **36** 2027

[16] Abdulla S, Diykh M, Laft R L, Saleh K and Deo R C 2019 Sleep EEG signal analysis based on correlation graph similarity coupled with an ensemble extreme machine learning algorithm *Expert Syst. Appl* **138** 112790

[17] Zhang X, Kou W, Eric I, Chang C, Gao H, Fan Y and Xu Y 2018 Sleep stage classification based on multi-level feature learning and recurrent neural networks via wearable device *Comput. Biol. Med.* **103** 71–81

[18] Fraiwan L, Lweesy K, Khasawneh N, Fraiwan M, Wenz H and Dickhaus H 2011 Time frequency analysis for automated sleep stage identification in fullterm and preterm neonates *J. Med. Syst.* **35** 693–702
[19] Fraiwan L and Lweesy K 2014 Newborn sleep stage identification using multiscale entropy *2014 Middle East Conf. on Biomed. Eng.* pp 361–4
[20] Koolen N, Oberdorfer L, Rona Z, Giordano V, Werther T, Klebermass-Schrehof K, Stevenson N and Vanhatalo S 2017 Automated classification of neonatal sleep states using EEG *Clin. Neurophysiol.* **128** 1100–8
[21] Čić M, Šoda J and Bonković M 2013 Automatic classification of infant sleep based on instantaneous frequencies in a single-channel EEG signal *Comput. Biol. Med.* **43** 2110–7
[22] Werth J, Serteyn A, Andriessen P, Aarts R M and Long X 2019 Automated preterm infant sleep staging using capacitive electrocardiography *Physiol. Meas.* **40** 55003
[23] Ansari A H, De Wel O, Pillay K, Dereymaeker A, Jansen K, Van Huffel S, Naulaers G and De Vos M 2020 A convolutional neural network outperforming state-of-the-art sleep staging algorithms for both preterm and term infants *J. Neural Eng.* **17** 16028
[24] Boostani R, Karimzadeh F and Nami M 2017 A comparative review on sleep stage classification methods in patients and healthy individuals *Comput. Methods Programs Biomed.* **140** 77–91
[25] De Wel O, Lavanga M, Dorado A C, Jansen K, Dereymaeker A, Naulaers G and Van Huffel S 2017 Complexity analysis of neonatal EEG using multiscale entropy: applications in brain maturation and sleep stage classification *Entropy* **19** 516
[26] Ghimatgar H, Kazemi K, Helfroush M S and Aarabi A 2018 An improved feature selection algorithm based on graph clustering and ant colony optimization *Knowledge-Based Syst.* **159** 270–85 https://doi.org/10.1016/j.knosys.2018.06.025
[27] Tsinalis O, Matthews P M, Guo Y and Zafeiriou S 2016 Automatic sleep stage scoring with single-channel EEG using convolutional neural networks (arXiv:1610.01683)
[28] Supratak A, Dong H, Wu C and Guo Y 2017 DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG *IEEE Trans.Neural Syst. Rehabil. Eng.* **25** 1998–2008
[29] Phan H, Andreotti F, Cooray N, Chén O Y and De Vos M 2019 SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging *IEEE Trans. Neural Syst. Rehabil. Eng.* **27** 400–10
[30] Pan S-T, Kuo C-E, Zeng J-H and Liang S-F 2012 A transition-constrained discrete hidden Markov model for automatic sleep staging *Biomed. Eng. Online* **11** 52
[31] Jiang D, Lu Y, Yu M A and Yuanyuan W 2019 Robust sleep stage classification with single-channel EEG signals using multimodal decomposition and HMM-based refinement *Expert Syst. Appl.* **121** 188–203
[32] Grigg-Damberger M, Gozal D, Marcus C L, Quan S F, Rosen C L, Chervin R D, Wise M, Picchietti D L, Sheldon S H and Iber C 2007 The visual scoring of sleep and arousal in infants and children *J. Clin. Sleep Med.* **3** 201–40
[33] Aarabi A, Wallois F and Grebe R 2006 Automated neonatal seizure detection: A multistage classification system through feature selection based on relevance and redundancy analysis *Clin. Neurophysiol.* **117** 328–40
[34] Aarabi A, Grebe R and Wallois F 2007 A multistage knowledge-based system for EEG seizure detection in newborn infants *Clin. Neurophysiol.* **118** 2781–97
[35] Gharbali A A, Najdi S and Fonseca J M 2018 Investigating the contribution of distance-based features to automatic sleep stage classification *Comput. Biol. Med.* **96** 8–23
[36] Steven M K 1988 *Modern Spectral Estimation: Theory and Application*
[37] Koolen N, Jansen K, Vervisch J, Matic V, De Vos M, Naulaers G and Van Huffel S 2014 Line length as a robust method to detect high-activity events: automated burst detection in premature EEG recordings *Clin. Neurophysiol.* **125** 1985–94
[38] Esteller R, Echauz J, Tcheng T, Litt B and Pless B 2001 Line length: an efficient feature for seizure onset detection *2001. Proc. of the 23rd Annual Int. Conf. of the IEEE Eng. Med. Biol. Soc.* pp 1707–10
[39] Koolen N, Dereymaeker A, Räsänen O, Jansen K, Vervisch J, Matic V, De Vos M, Naulaers G, Van Huffel S and Vanhatalo S 2015 Data-driven metric representing the maturation of preterm EEG *2015 37th Annual Int. Conf. of the IEEE Eng. Med. Biol. Soc.* pp 1492–5
[40] Kern S J 2017 Automatic sleep stage classification using convolutional neural networks with long short-term memory *Master's Thesis* Radboud University
[41] Ghaemi A, Rashedi E, Pourrahimi A M, Kamandar M and Rahdari F 2017 Automatic channel selection in EEG signals for classification of left or right hand movement in Brain Computer Interfaces using improved binary gravitation search algorithm *Biomed. Signal Process. Control* **33** 109–18
[42] Khodabakhsh A, Ari I, Bakır M and Alagoz S M 2019 Forecasting multivariate time-series data using LSTM and MINI-BATCHES *The 7th Int. Conf. on Contemporary Issues in Data Science* (Springer) pp 121–9
[43] Dereymaeker A, Pillay K, Vervisch J, Van Huffel S, Naulaers G, Jansen K and De Vos M 2017 An automated quiet sleep detection approach in preterm infants as a gateway to assess brain maturation *Int. J. Neural Syst.* **27** 1750023
[44] Gerla V, Lhotska L, Krajca V and Paul K 2006 Multichannel analysis of the newborn EEG data *IEEE, ITAB, Int. Special Topics Conf. on Information Technology in Biomedicine* pp 1–6
[45] Gerla V, Bursa M, Lhotska L, Paul K and Krajca V 2007 Newborn sleep stage classification using hybrid evolutionary approach *Int. J. Bioelectromagn.* **9** 25–26
[46] Piryatinska A, Terdik G, Woyczynski W A, Loparo K A, Scher M S and Zlotnik A 2009 Automated detection of neonate EEG sleep stages *Comput. Methods Programs Biomed.* **95** 31–46
[47] Löfhede J, Thordstein M, Löfgren N, Flisberg A, Rosa-Zurera M, Kjellmer I and Lindecrantz K 2010 Automatic classification of background EEG activity in healthy and sick neonates *J. Neural Eng.* **7** 16007
[48] G L H and C T L 1993 Prognostic value of background patterns in the neonatal EEG *J. Clin. Neurophysiol.* **10** 323–52
[49] Malafeev A, Laptev D, Bauer S, Omlin X, Wierzbicka A, Wichniak A, Jernajczyk W, Riener R, Buhmann J and Achermann P 2018 Automatic human sleep stage scoring using deep neural networks *Front. Neurosci.* **12** 781