Journal of Neural Engineering

PAPER • **OPEN ACCESS**

Automated EEG sleep staging in the term-age baby using a generative modelling approach

To cite this article: Kirubin Pillay *et al* 2018 *J. Neural Eng.* **15** 036004

View the article online for updates and enhancements.

Recent citations

- Neonatal EEG sleep stage classification based on deep learning and HMM
Hojat Ghimatgar *et al*
- A convolutional neural network outperforming state-of-the-art sleep staging algorithms for both preterm and term infants
Amir H Ansari *et al*
- A perinatal stress calculator for the neonatal intensive care unit: an unobtrusive approach
M Lavanga *et al*

OPEN ACCESS
IOP Publishing
Journal of Neural Engineering
J. Neural Eng. **15** (2018) 036004 (13pp)
https://doi.org/10.1088/1741-2552/aaab73

Automated EEG sleep staging in the term-age baby using a generative modelling approach

Kirubin Pillay[1], Anneleen Dereymaeker[2], Katrien Jansen[2,3], Gunnar Naulaers[2], Sabine Van Huffel[4,5] and Maarten De Vos[1]

[1] Department of Engineering Science, Institute of Biomedical Engineering (IBME), University of Oxford, Oxford, United Kingdom
[2] Department of Development and Regeneration, University Hospitals Leuven, Neonatal Intensive Care Unit, KU Leuven (University of Leuven), Leuven, Belgium
[3] Department of Development and Regeneration, University Hospitals Leuven, Child Neurology, KU Leuven (University of Leuven), Leuven, Belgium
[4] Division Stadius, Department of Electrical Engineering-ESAT, KU Leuven (University of Leuven), Leuven, Belgium
[5] IMEC, Leuven, Belgium

E-mail: kirubin.pillay@eng.ox.ac.uk

Received 13 December 2017
Accepted for publication 30 January 2018
Published 27 February 2018

Abstract

Objective. We develop a method for automated four-state sleep classification of preterm and term-born babies at term-age of 38–40 weeks postmenstrual age (the age since the last menstrual cycle of the mother) using multichannel electroencephalogram (EEG) recordings. At this critical age, EEG differentiates from broader quiet sleep (QS) and active sleep (AS) stages to four, more complex states, and the quality and timing of this differentiation is indicative of the level of brain development. However, existing methods for automated sleep classification remain focussed only on QS and AS sleep classification. *Approach.* EEG features were calculated from 16 EEG recordings, in 30 s epochs, and personalized feature scaling used to correct for some of the inter-recording variability, by standardizing each recording's feature data using its mean and standard deviation. Hidden Markov models (HMMs) and Gaussian mixture models (GMMs) were trained, with the HMM incorporating knowledge of the sleep state transition probabilities. Performance of the GMM and HMM (with and without scaling) were compared, and Cohen's kappa agreement calculated between the estimates and clinicians' visual labels. *Main results.* For four-state classification, the HMM proved superior to the GMM. With the inclusion of personalized feature scaling, mean kappa (±standard deviation) was 0.62 (±0.16) compared to the GMM value of 0.55 (±0.15). Without feature scaling, kappas for the HMM and GMM dropped to 0.56 (±0.18) and 0.51 (±0.15), respectively. *Significance.* This is the first study to present a successful method for the automated staging of four states in term-age sleep using multichannel EEG. Results suggested a benefit in incorporating transition information using an HMM, and correcting for inter-recording variability through personalized feature scaling. Determining the timing and quality of these states are indicative of developmental delays in both preterm and term-born babies that may lead to learning problems by school age.

Original content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Keywords: EEG, hidden markov models, Gaussian mixture models, automated sleep staging, generative models, preterm newborn, term newborn

(Some figures may appear in colour only in the online journal)

1. Introduction

In the Neonatal Intensive Care Unit (NICU), both term-born and preterm infants are treated. Preterm infants are born at <37 weeks postmenstrual age (PMA, the age since the last menstrual cycle of the mother) and are the largest group of patients treated in the NICU. These patients are mostly discharged at a PMA of >36 weeks. Preterms are of heightened vulnerability and are most at risk of disturbances to normal brain development during this period of acute illness, that may only become apparent by school-age [1–4]. On the other hand, term-born infants (PMA of >37 weeks) are admitted because of congenital anomalies or acute perinatal disease. Both cohorts of babies spend a large proportion of their early life in sleep, whereby the brain continues to build complex cortical pathways leading to memory and enhanced cognition [5, 6]. Better understanding of the precise nature and transitions of the different sleep states can help identify potential developmental delays early, which result from these various cognitive stresses [7–11].

Neonatal sleep from 28 weeks PMA is broadly separated into quiet sleep (QS), active sleep (AS) and Wake. AS is also known as rapid-eye movement (REM) sleep, as periods of REM are typical during this state, as well as variable breathing, heart rate, and motion. QS (non-REM sleep), on the other hand, is characterized by reduced eye movement, regular breathing, and a steadier heart rate [5, 12–14]. These sleep states are also identifiable (and well documented) using multichannel electroencephalogram (EEG) recordings, which can be non-invasively measured in the NICU [5, 12, 13, 15, 16]. There is also a further indeterminate state defined as having non-EEG characteristics of AS and EEG characteristics of QS, or vice versa, which often occurs when transitioning from one sleep state to the next [12, 17–21].

From preterm to beyond term-age, the proportions and characteristics of these EEG sleep states change, as illustrated in figure 1(a). It is only at the onset of term-age (36–37 weeks PMA), that differentiation to further complex sleep patterns can be defined from the EEG, based on particular changes in amplitude and frequency behaviours [5, 12, 15, 16, 22–25]. AS separates into two forms, denoted AS I and AS II, respectively. AS I is defined by mixed frequency EEG behaviour with variable amplitudes as shown in figure 1(b) and AS II consists of a low-voltage irregular (LVI) pattern with consistently lower amplitudes than AS I (figure 1(c)). Before term-age, QS has a distinct pattern of alternating bursts and suppression called Tracé Discontinu (TD), but by term age, this state has matured into Tracé Alternant (TA), where the suppression periods now have a heightened amplitude (figure 1(d)). There is also a further QS state of high voltage slow-wave (HVS), with higher amplitude, lower frequency characteristics than in AS (figure 1(e)).

Analysing the EEG for both preterm and term-born babies, at term-age, can provide a quantitative measure of the quality of differentiation into the further sub-states, with the absence or prematurity of certain states providing an early indication of delayed sleep development. For instance, the increasing proportion of HVS and decrease of TA (within the same QS state) remains an important indication of normal maturation beyond 38 weeks PMA [25].

However, to achieve this, there is a need to efficiently classify these sleep states automatically, as visual labelling requires expert training and is time consuming in the clinical setting [12]. Achieving automated sleep classification can also improve the quality and timing of neonatal care, intervention, and treatment, while minimizing disturbances to the baby's sleep [22, 26, 27].

Few methods exist for automated sleep classification at term-age. Turnbull *et al* uses discrete wavelet transforms (DWT) and EEG features for the specific detection of TA [24], but not the other potentially important sleep states. Gerla *et al* introduces a 'hybrid evolutionary approach' combining features from full polysomnographic (PSG) signals (EEG as well as electrocardiogram (ECG) for heart rate, measures of respiratory rate, body movement, and eye movements) [28]. However, often only EEG (and ECG) are consistently recorded and therefore developing an approach focussed on EEG alone would be more applicable in the NICU.

More recently, De Wel *et al* developed a supervised approach for the detection of QS and non-QS across a wide range of preterm and term ages, using an LS-SVM classifier and multiscale entropy features [29]. We have also developed an unsupervised method for the successful detection of QS and non-QS periods across a similar age range, called cluster-based adaptive sleep staging (CLASS) [14]. Although successful, these existing approaches cannot classify all four EEG states at term-age. This motivated the development of an approach for automated sleep classification of term-age EEG to achieve precisely this task.

The visual scoring of sleep usually incorporates knowledge of the previous sleep state to influence the decision on the current state, with transitions between certain states physiologically more likely than others [12, 17]. This information can be incorporated using a transition-based generative model, namely the Hidden Markov model (HMM). Often, the observed feature distributions underpinning the HMMs are modelled using Gaussian mixture models (GMMs), as they can fit arbitrarily complex (multi-modal) distributions.

In this paper, we present an HMM classifier to perform automated sleep staging of EEG, using GMMs to model the observed EEG feature distributions. To assess the importance of accounting for sleep-state transitions, we also compare sleep classification by the HMM to classification by the GMM alone. We first present a binary classification, performing broader QS and non-QS state detection as a benchmark to compare these models with existing approaches,

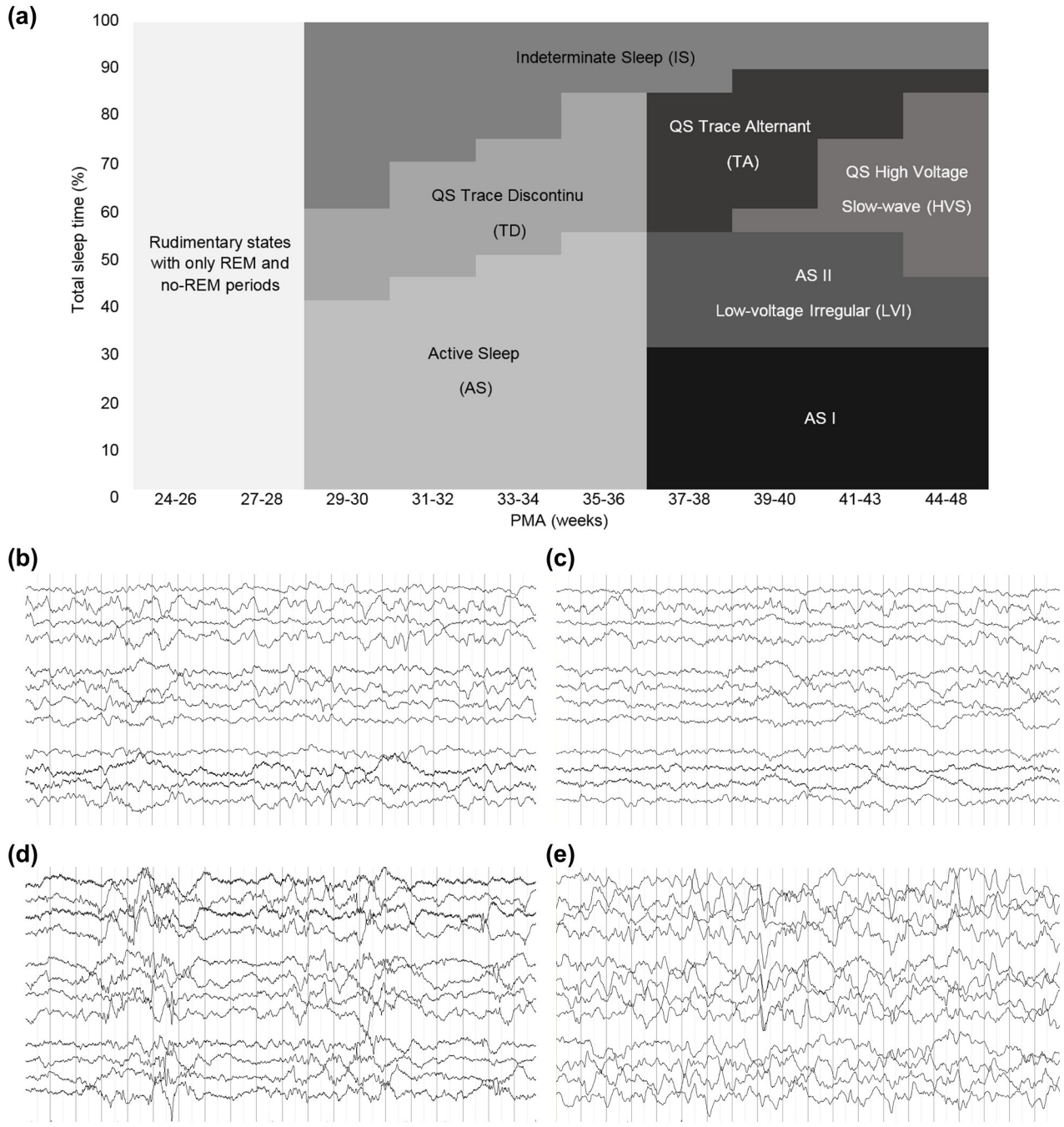


Figure 1. (a) The evolution of the EEG sleep states from preterm to term ages, highlighting the differentiation of QS and AS into four sub-states at 36–37 weeks PMA. (b) 30 s 12-channel montage EEG recording at term-age, of active sleep I (AS I) consisting of mixed frequencies and variable amplitudes. (c) active sleep II (AS II) state, showing a low-voltage irregular (LVI) behaviour characterized by consistently lower frequencies than AS I. (d) TA, consisting of bursts of high amplitude activity separated by suppressed periods known as inter-burst intervals. (e) High-voltage slow-wave sleep (HVS), consisting of typically higher voltage behaviour than the AS I and AS II states with less regular amplitude fluctuations over time.

before developing the more complex models for four-state classification.

2. Methods

2.1. Data acquisition and EEG recordings

Data was recorded at the NICU of the University Hospitals, Leuven, Belgium and approved by the ethics committee of the University Hospitals, Leuven. Newborns were enrolled after prior consent by the parents. A total of 16 preterm and term-born newborns of 27–41 weeks gestational age (their age at birth) were pre-selected based on a 'normal' developmental outcome at nine and 24 months of age. Normal developmental outcome was defined by:

1. Bayley scores of infant mental and motor development >85,
2. No use of any sedatives or anti-epileptic medication during EEG recording,

3. The absence of any severe cerebral lesions, as assessed by cranial ultrasound.

One EEG recording was measured from each patient at term age, resulting in a total of 16 EEG recordings aged 38–42 weeks PMA. Mean recording length was 7 h 55 min (range 1 h 58 min–17 h 50 min). All EEG recordings were recorded with nine electrodes (Fp1, Fp2, C3, C4, T3, T4, O1, O2 and a reference Cz) at 250 Hz sampling frequency, using the standard 10–20 electrode system. In addition to EEG monitoring, these infants had a full PSG recording with non-cerebral measures including ECG, oxygen saturation, piezo-electric belts (measuring abdominal and thoracic respiratory effort), two electro-oculograms, and a nasal thermistor (for airflow monitoring).

2.2. Visual sleep scoring

EEG segments were visually labelled by two trained raters. The primary rater (AD) identified the start and end periods of the EEG for two states of AS (AS I and AS II), Wake, two states of QS (HVS and TA), indeterminate sleep, and regions of artefact where a clear sleep state could not be identified. The second rater (KJ) reviewed the labels of the primary rater, and identified those that were not agreed on. These 'dubious' regions, in addition to artefacts and indeterminate periods, were removed from further analysis.

Sleep states were defined by reviewing both EEG and non-cerebral characteristics, considering the presence of REM, cardiorespiratory regularity, and body movements. AS II and Wake both consist of an LVI pattern and were difficult to distinguish from EEG alone, with non-cerebral characteristics required (eyes open/closed and changes in body movement) to separate the two states [12, 17]. Therefore, for this EEG-based study, these were subsequently merged into a single LVI state, with the final sleep states for classification defined as AS I, HVS, TA and LVI.

2.3. EEG pre-processing and feature extraction

EEG recordings were processed at their original sampling frequency of 250 Hz, and bandpass filtered in the range 0.5–40 Hz, using 250-order FIR low-pass and high-pass filters.

A set of EEG features were calculated as potential candidates for classifying term-age sleep, the majority of which have been well defined in previous EEG studies on preterm and term-age sleep staging, adult sleep staging and newborn seizure detection. The full list is presented in table 1, including references to existing EEG applications and definitions, where relevant.

Features were obtained from the time-domain and Fourier-domain, in the frequency bands delta (0.5–3 Hz), theta (3–8 Hz), alpha (8–12 Hz) and beta (12–30 Hz), including measures of EEG complexity, such as fractal dimension [23–32] and Hjorth parameters [30, 33–36]. To account for the non-stationarity of EEG, we also computed features from wavelet [30, 34–36] and empirical mode decompositions (EMDs) [30, 37].

In addition to these classical metrics, we included recent Multiscale Entropy (MSE) measures that have proven successful for sleep staging in preterm and term-age EEG [29].

The method of line length (LL) provided a robust measure for EEG suppression [38–40], while a measure of peak-to-peak EEG amplitude, called rEEG, was included as it has previously identified burst and suppression periods in newborn EEG with great success [41]. For further details on these MSE, LL and rEEG features, and calculation of the Fourier and wavelet transforms, we direct the reader to the Appendix.

Previous sleep staging studies in preterms have used epoch lengths as long as 2.5 min [39], with certain features, such as MSE, proving to be unreliable at window lengths as short as 10 s [29]. However, excessively long epoch lengths can miss brief state transitions, while lengths still need to be sufficiently long enough to capture the burst-suppression cycles observed during TA (where suppression lengths are typically $\leqslant$6 s [25]). We therefore chose a trade-off epoch length of 30 s, with an overlap of 5 s to account for potential sleep state transitions at the epoch boundaries. Features were calculated for each epoch and EEG channel with the median across channels taken to minimize the effects of channel-specific deviations. A total of 112 features were calculated per epoch.

2.4. Personalized feature scaling

When assessing the feature distributions across states and recordings, the scale and range of the data varied noticeably from one recording to the next. Figure 2(a) illustrates this variability between four patient recordings for the feature: SD of theta band amplitudes, across the four labelled sleep states.

To partially correct for inter-subject variability in the EEG, due to differing electrode placements and baseline changes, 'personalized feature scaling' has been proposed to convert the extracted features from different subjects to the same scale [53]. We apply this idea here to correct for inter-recording variability, by standardizing each recording's features with its own mean and standard deviation. Figure 2(b) shows the same feature and recordings as in figure 2(a), after scaling, showing a reduced variability across recordings.

Each recording was individually scaled before performing subsequent feature selection and classification.

2.5. Feature selection

Feature selection is particularly important when fitting GMM distributions, as too many features can prevent convergence or an optimal fit (the 'curse of dimensionality'). With a large set of 112 features calculated, high correlations between some features were inevitable. While the covariance of the GMMs can account for some correlation, it could still be problematic if too high.

Several common feature selection methods exist, which aim to maximize the correlation between the features and the sleep state, while most also attempt to remove redundancies between highly correlated features. We chose the method of minimum redundancy maximum relevance (mRMR), which can operate on continuous features by using spearman rank correlation coefficient as the selection criteria, instead of the previously proposed mutual information [54]. The mRMR method finds a trade-off between those features that separate best between sleep states (using the visual labels), and those

Table 1. Features calculated from each EEG epoch, including time-domain and frequency domain measures. References to previous applications in EEG (predominantly for sleep staging) are also provided.

Feature category	Feature description	References
Time domain	SD, mean absolute amplitudes	[14, 31, 34, 36, 42–44]
	Max. Absolute amplitude, and sum of first and second derivatives	[14, 42]
	Max–min difference of amplitudes	[14, 42]
	Skewness and Kurtosis	[23, 31, 33, 34, 36, 44]
	Hjorth activity, complexity, mobility	[30, 33–36, 52]
	ApEn, SampEn, and MSE (scales 1–10) SD of MSE values	[30, 34, 35, 45]
	Area under multiscale curve	[29]
	Average slope of multiscale curve (scales 1–5)	[29]
	Average slope of multiscale curve (scales 6–20)	[29]
	Max MSE value	[29]
	Katz and Higuchi fractal dimensions ($k_{max} = 20, 40, 60$)	[23, 30–32]
	Zero crossing rate	[30, 34–36, 46]
	Coastline distance	[33]
	Mean and SD TKEO values	[34, 35]
	90% peak value, width, prominence (SD)	[33]
	LL suppression value (supp)	[38–40]
	Mean, median, LM (5th percentile), HM (95th percentile), SD, IQR, and skewness of the rEEG	[39, 41]
Fourier transform	SD and mean absolute amplitudes in delta, theta, alpha, beta frequency bands	[10]
	Hilbert median envelope amplitude, delta, theta	
Wavelet decomposition (debauchies-4 wavelet)	Mean absolute value and SD of amplitudes in D3, D4, D5, A5 bands	[30, 34]
	Ratio of absolute mean values in adjacent sub bands	
	Sum of squared coefficients in D3, D4, D5, A5 bands	[30, 52]
EMD	Variation coefficients for IMF1–IMF6	[30, 37]
	Fluctuation index of IMF1–IMF6	[30, 37]
	Mean absolute ratio of each pair of successive IMFs	
	Hilbert median instantaneous frequency, IMF1–IMF6	[47, 48]
Frequency domain	Full band power (delta-beta range)	[32, 35, 36, 49, 50]
	Mean frequency	[42, 44]
	Spectral roll-off SD, centroid SD, flux SD	[46]
	Power spectral entropy	[23, 30, 31, 34–36]
	90% spectral edge frequency	[23, 31, 35, 36, 43, 46]
	Relative band power in delta, theta, alpha, beta bands	[23, 39, 51]
	Mean frequency in delta, theta, alpha, beta bands	[44]
	EEG spectral beta/alpha ratio	[9]
	Hilbert median instantaneous frequency, delta, theta bands	

SD: standard deviation, ApEn: approximate entropy, SampEn: sample entropy, MSE: multiscale entropy, EMD: empirical mode decomposition, IMF: intrinsic mode function, TKEO: Teager–Kaiser energy operator, LL: line length. Delta: 0.5–3 Hz, theta: 3–8 Hz, alpha: 8–12 Hz, beta: 12–30 Hz.

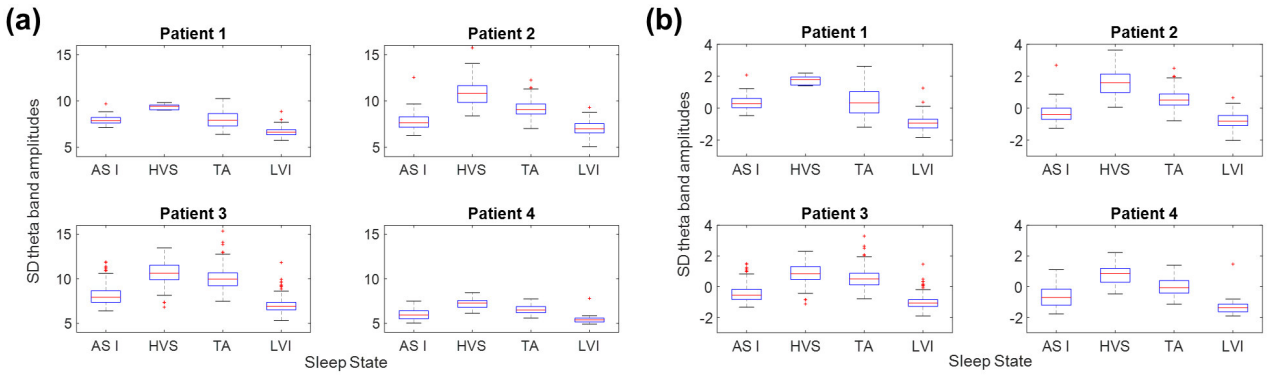


Figure 2. Illustration of personalized feature scaling on each patient recording, to partially correct for inter-recording variability by standardizing each recording with its mean and standard deviation. (a) The original distributions of the feature: SD of theta band amplitudes, across the four sleep states and for four different patient recordings. (b) The corresponding distributions from (a) after personalized feature scaling.

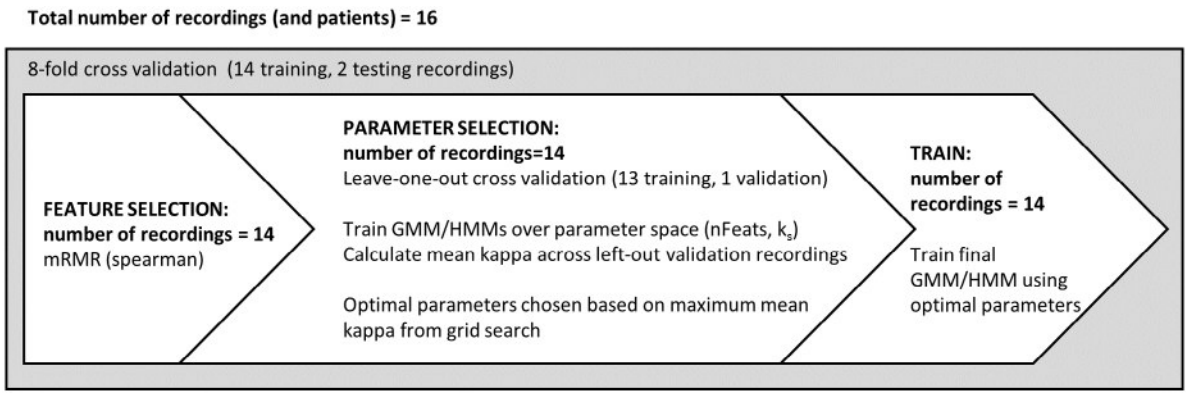


Figure 3. Overview of the sleep classification strategy. 8-fold cross validation is used to separate training and testing recordings for training the GMM and HMM models and assessing classifier performance. Within each training fold, feature selection is applied, before a further leave-one-out cross validation to identify the optimal parameters of the model (number of selected features (nFeats) and number of GMM components for each sleep state (k_s)).

that do not correlate strongly with each other. The output is a full re-ranked list of the features.

2.6. Sleep classification

We denote each recording for classification by a feature matrix, **X**, consisting of m selected features by n epochs, or observations (also referred to as the observation sequence). Each observation in **X** is represented by an observation vector, **x** (of length m). Models were fitted on a training set of the recordings before classification on the test set. The precise methods for training and testing on the data set, and the tuning of parameters, are detailed in section 2.7.

2.6.1. Classification using GMMs. Each observation vector, **x**, was independently classified to a sleep state, s, by generating a posterior probability using a uniquely fitted likelihood distribution, p(**x**|s), for each state.

GMMs are commonly used to define p(**x**|s), and can account for some correlation between features (when choosing a full covariance structure). A GMM was fitted to each sleep state distribution using the labelled training data, by Expectation-Maximization [55]. For each sleep state, p(**x**|s) is defined by:

$$\mathrm{p}\left(\mathbf{x}|\mathrm{s}\right) = \sum_{i=1}^{k_\mathrm{s}} c_{i\mathrm{s}} \mathcal{N}_{i\mathrm{s}}\left(\boldsymbol{\mu}_{i\mathrm{s}}, \boldsymbol{\Sigma}_{i\mathrm{s}}\right). \quad (1)$$

$\mathcal{N}_{i\mathrm{s}}$ denotes the ith Gaussian component of the sleep-state-specific mixture model. Component weights are given by $c_{i\mathrm{s}}$, and $\boldsymbol{\mu}_{i\mathrm{s}}$ and $\boldsymbol{\Sigma}_{i\mathrm{s}}$ denote the mean and covariance of each component, respectively. The number of Gaussian components for each sleep state GMM is denoted by k_s. Using Bayes' rule, a posterior probability, p(s|**x**), is obtained for each sleep state, weighted by the prior probability of the state, p(s), and the observation likelihood obtained from equation (1):

$$\mathrm{p}\left(\mathrm{s}|\mathbf{x}\right) = \frac{\mathrm{p}\left(\mathbf{x}|\mathrm{s}\right)\mathrm{p}\left(\mathrm{s}\right)}{\sum_\mathrm{s} \mathrm{p}\left(\mathbf{x}|\mathrm{s}\right)\mathrm{p}\left(\mathrm{s}\right)}. \quad (2)$$

The priors, p(s), were estimated directly from the counts of the visual labels in the training set, and incorporating this prior knowledge accounted for some sleep states that occurred more commonly than others. The maximum value of p(s|**x**) provided the final sleep state estimate.

2.6.2. Classification using HMMs. The HMM was implemented by extending the GMM model with the addition of sleep state transition probabilities. HMMs typically assume a first-order Markov assumption, where the next sleep state (of the next observation) depends only on the current state. These states form a hidden (unobserved) layer, and movement between states are defined using a transition matrix, **T**, which represents the transition probability from the current sleep state to the next. Transitions to a hidden state, s, then 'generates' an observable vector **x** from the observation distribution, p(**x**|s). By knowing the full observation sequence, **X** (and **T**) this procedure can be reversed and the most probable sequence of hidden sleep states inferred.

The GMM observation distributions, p(**x**|s), were similarly calculated from equation (1), and the same priors, p(s), used for defining the initial sleep state. **T** was estimated by directly counting the number of sleep transitions from the training labels.

Classifying the full sequence, **X**, was performed by the Viterbi algorithm [56], which estimates the most likely sequence of sleep states that could have generated **X**. We calculated this in the log-domain to prevent computational underflow, a problem when dealing with particularly long sequences.

2.7. Training and testing the models

The detailed steps of the testing-training procedure are shown in figure 3. Both the GMM and HMM were trained concurrently. We applied an 8-fold cross validation to first separate into 14 training and 2 test recordings. Within each training fold, feature selection was performed followed by a leave-one-out cross validation to select the optimal parameters (further removing a recording from the training set for validation). Parameters requiring tuning were the number of selected features (nFeats) and the number of components for each sleep-state GMM (k_s). A grid search was used, repeatedly training

Table 2. Classification results for QS and non-QS sleep classification. Comparisons are made between GMMs and HMMs, with and without feature scaling, as well as the CLASS QS detection algorithm.

Model	Mean kappa (SD)	Mean % acc. (SD)	Mean % sens. (SD)	Mean % spec (SD)	p_{boot}
HMM	0.89 (0.063)	95 (3.1)	94 (4.1)	95 (3.6)	—
GMM	0.85 (0.11)	92 (5.5)	94 (5.9)	91 (6.6)	0.19
HMM (no scaling)	0.81 (0.16)	91 (8.6)	89 (16)	93 (9.0)	0.37
GMM (no scaling)	0.78 (0.21)	89 (11)	87 (21)	92 (9.2)	0.18
CLASS	0.68 (0.22)	84 (14)	72 (22)	96 (4.7)	0.00

acc: accuracy, sens: sensitivity, spec: specificity. SD: standard deviation. p_{boot} = proportion of bootstrapped recordings where model performs better than the HMM (with scaling).

GMMs and HMMs for each parameter combination and assessing mean performance across the left-out validation recordings using Cohen's kappa agreement between the clinicians' sleep labels and the model's estimates. The maximum mean kappa across the parameter space identified the optimal parameters. In the final step, the selected features and optimal parameters were used to train a GMM and HMM on the full training fold and performance assessed on the test recordings.

2.8. Assessing performance

Metrics used to assess and compare the performance of the models, were kappa, accuracy, sensitivity and specificity. In the first case, we trained HMM and GMM models for two-state sleep classification of the broader QS and non-QS states, as a benchmark, to assess the performance of these generative models when compared to existing methods. We directly compared these models to the CLASS QS detection method as it could be confidently tested on the same dataset used in this study. The personalized scaling of recordings, while used in adult EEG, has not been previously explored for preterm EEG analysis. We therefore also trained the GMM and HMM models, with and without scaling for comparison, to better understand the importance of this step.

In the second case, we performed the four-state sleep classification. As in the two-state case, models were trained with and without scaling and compared.

3. Results

3.1. Two-state classification performance

For two-state classification, a mean of 7.3 ($\pm$2.8) features across the 8 training folds were selected for the GMM, and a mean of 8.8 ($\pm$1.3) features for the HMM. We identified the most prominent features, as selected by mRMR, by determining an overall weighted ranking from the individual feature ranks across the training folds. As a (rounded) mean of 9 features were most selected by either model, we list these 9 most prominent features here, which were: Variation coefficient of IMF 1, Variation coefficient of IMF 2, EEG spectral beta/alpha ratio, SD of the MSE values, SD of the alpha band amplitudes, Variation coefficient of IMF 3, Ratio of absolute mean values of D4/D3, IQR of the rEEG, and SD of the theta band amplitudes.

Table 2 shows the overall test performance for both the GMM and HMM. These are compared to models trained without personalized feature scaling, and the CLASS algorithm. CLASS classifies QS periods as a whole 'event' rather than per epoch. Therefore, to allow for accurate comparison of performance, we converted the resulting identified periods into the corresponding 30 s epochs as classified by the GMM and HMM models. Mean results are shown with standard deviation in brackets. While the kappa values are interpreted differently depending on application, it is commonly regarded that values $<$0.4 are poor while values $>$0.75 are considered near-perfect [57, 58].

In terms of mean and SD, both models show consistently good results across the different metrics, performing similarly well with near-perfect mean kappa values for the GMM and HMM of 0.85 ($\pm$0.11) and 0.89 ($\pm$0.063) respectively, and accuracy, sensitivity and specificity values all $>$90%. When compared to the models without scaling, the mean kappa values are lower with higher standard deviations. Specificity remains very high across both scaled and unscaled models ($>$90%), with the drop in overall performance largely due to sensitivity.

Across all metrics, CLASS performs less well with mean kappa of 0.68 ($\pm$0.22) and struggles to accurately detect the QS periods in some cases. These results suggest that the GMM and HMM models (with scaling) are both excellent candidates for two-state classification, with exceptionally high and consistent performances.

Due to the similarities in some of the mean and SD values, we included an additional metric to better assess which models were more reliable. One would typically perform a statistical test, such as the paired *t*-test, however, due to the limited number of data points (only 16) and the ambiguity of the underlying distribution, such methods would be unreliable. We instead devised a method based on bootstrapping, comparing each model to the HMM (with scaling), based on the kappa values. We subtracted the HMM kappa results with the corresponding kappas of the model in question (between the same recordings). This resulted in 16 kappa differences, which were then bootstrapped (sampling with replacement) to 10 000 instances, a method first introduced by Efron [59]. From the bootstrapped set, we calculated the proportion of instances $<$0 (denoted p_{boot}). We can interpret p_{boot} as follows: what proportion of instances will the model in question perform better than the HMM? A p_{boot} $<$0.5 suggests the HMM is a more reliable choice as it would better classify the majority of recordings, while values $>$0.5 would suggest the opposite. These p_{boot} values are included in table 2 and are all $<$0.5, suggesting that the HMM is always a more reliable model choice.

Table 3. The 17 most prominent features selected across the 8 training cross validation folds, for four-state classification. These are identified from an overall weighted ranking based on the individual feature ranks across the training folds.

Feature name
Hilbert median instantaneous frequency, IMF1
Fluctuation index of IMF4
Mean absolute ratio, IMF3/IMF2
Mean absolute amplitudes, beta
Mean of absolute values in sub-band, D3
Fluctuation index of IMF1
Hilbert median instantaneous frequency, IMF4
Average slope of multiscale curve (scales 6–20)
SD of the amplitudes in sub-band, D3
LM rEEG
Relative bandpower in alpha
Hilbert median envelope amplitude, theta
Max. MSE value
Area under multiscale curve
SD of the beta band amplitudes
MSE, scale 9
SD of the amplitudes in sub-band, D5

3.2. Four-state classification performance

In the four-state case, a mean of 17 (±3.1) features were selected for the GMM, and 16 (±2.1) features for the HMM. More features were required for each model, than in the two-state case, and the 17 most prominent features appearing across the training folds are listed in table 3.

Figure 4 shows the training performance curves over feature number (up to 30 features) for both the GMM (figure 4(a)) and HMM (figure 4(b)). Grey curves show the performance at each training fold, overlaid with the mean performance curve across all folds. We see an overall higher kappa in the HMM, compared to the GMM, with both models showing a smooth improvement with the addition of more features, peaking at 15–20 features. Performance plateaued beyond this number.

Figure 5(a) shows an example of the sleep classification hypnograms by the clinicians, GMM and HMM models for a single test recording, and figure 5(b) shows the average transition matrix across the training folds. From these figures, we see that a high probability of self-transitions (>0.9) provides a smoother, improved hypnogram estimate (more similar to the clinicians') than the GMM model, where more regular, spurious transitions are observed. Figures 5(c) and 5(d) show the combined confusion matrices of the GMM and HMM, respectively, compared to the true (clinicians') labels. For the HMM, a much higher proportion of sleep epochs are correctly classified across all four sleep states than the GMM. The majority of misclassifications in both models occur between each pair of QS (HVS and TA), and non-QS (AS I and LVI) states.

Table 4 shows the overall test performance when compared to the models without scaling. Accuracy, sensitivity and specificity measures are determined as the state-averaged accuracy, sensitivity and specificity across the four sleep states. Values for p_{boot} are also shown.

As alluded to by the confusion matrices, overall kappa confirms that the HMM clearly outperforms the GMM, achieving a higher value of 0.62 (±0.16) compared to the GMM of 0.55 (±0.15). Without scaling, HMM kappa falls to 0.56 (±0.18). However, this is still an improvement on the GMM, which shows the poorest performance of all the models, with a kappa of 0.51 (±0.15).

4. Discussion

The HMM has proven a successful classifier for sleep staging in adult EEG [30], but to the best of our knowledge, this is the first study that uses such methods successfully for four-state sleep classification of term-age babies. This is a critical age at which differentiation into further sub-states are identified, with the quality and rate of maturation of these states providing a potential biomarker for identifying abnormal development, in both preterm and term-born babies. We focussed on EEG alone to successfully classify these states, as this is a consistently measured signal in these babies, and the HMM further incorporates knowledge of sleep-state transitions, as is often done during visual labelling, which improves overall performance.

Features derived from wavelet decompositions and EMDs were most prominently selected by mRMR, suggesting a benefit of accounting for the non-stationary nature of the EEG, which is not considered by more traditional Fourier transform-based metrics (particularly when long epochs lengths are considered and the signal can no longer be approximated as stationary). Furthermore, such features derived from decomposed EEG signals, reflects existing knowledge that specific sleep states are defined by distinct behaviours that uniquely vary across the frequency bands [25].

A greater number of features were selected in the four-state case, compared to the two-state case and this is also indicative of the greater complexity required by the models to distinguish between the four, more subtly different states, as opposed to simply QS and non-QS.

The recently introduced MSE-based features and the lower margin (LM) rEEG measures were regularly selected in four-state classification. These features have already shown promise for identifying suppression and maturation state from EEG at preterm and term ages [29, 41] and seem to identify specific sleep states here, due to the varying complexity and burst-suppression behaviours of the sleep states (such as during TA). With this said, we stress that with such a large set of features extracted, the choice of feature selection approach plays a big part. Due to the potential for high redundancy between features, we chose mRMR to simultaneously rank those features strongly discriminating for sleep state, while ensuring a limited redundancy between them, based on spearman correlation coefficients. However, the use of different metrics for feature selection may well identify a very different optimal feature subset with a similar classification performance achieved.

Both the GMM and HMM classifiers show a consistently excellent two-state performance even with a model, such as a

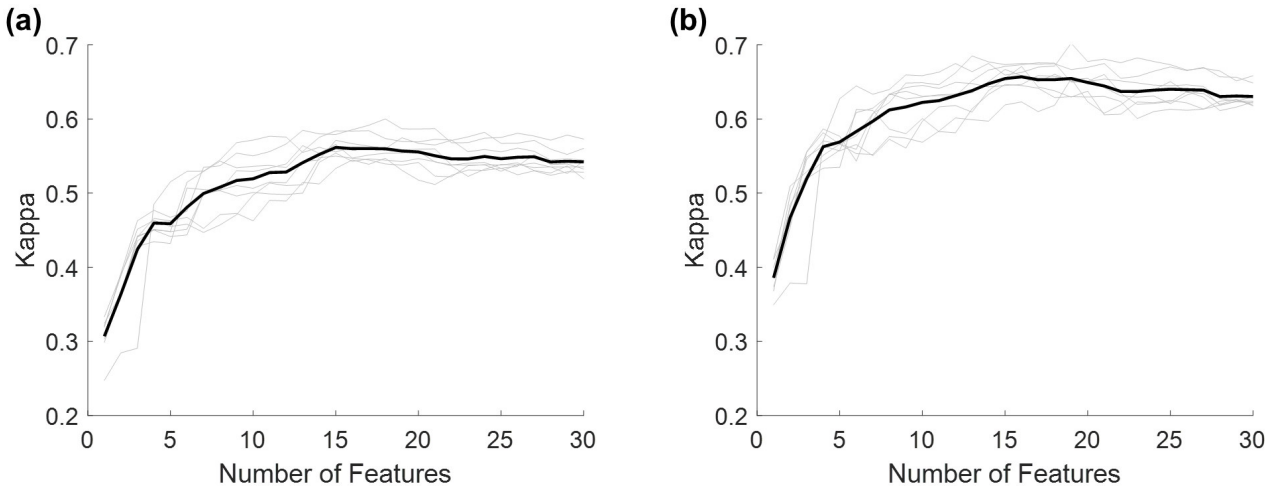


Figure 4. (a) Training performance curve for the GMM model as the number of features are increased (up to 30). Grey curves denote performance at each training fold, with the black curve denoting the mean performance across folds. (b) Corresponding training performance curve for the HMM.

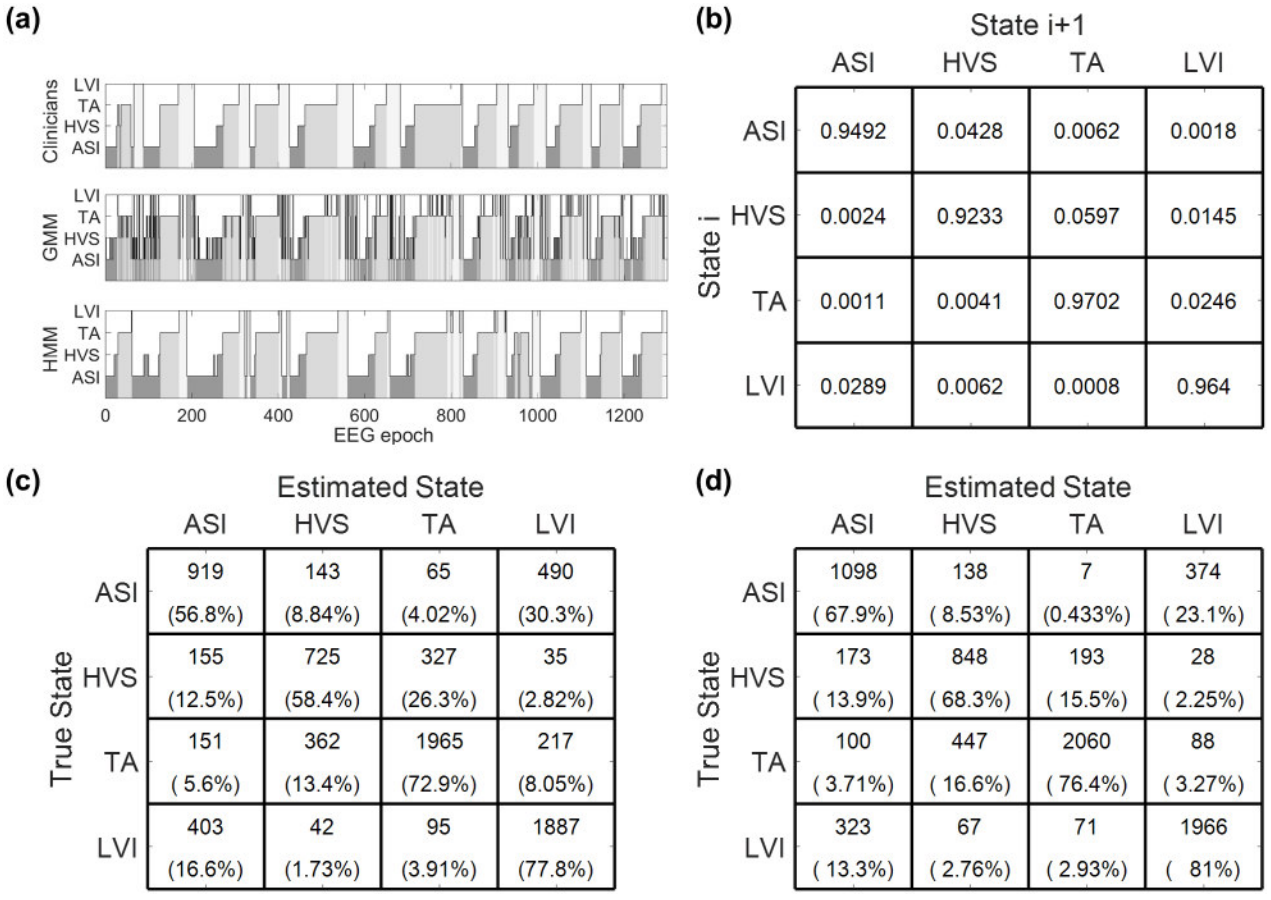


Figure 5. (a) Example hypnogram of four-state sleep classification on a single recording, by the clinicians (top), GMM model (middle) and HMM (bottom). (b) Average HMM transition matrix across training folds. (c) Confusion matrix comparing total number of correctly classified epochs between the true (clinician) states and the GMM estimates for active sleep I (ASI), high-voltage slow-wave (HVS), Tracé Alternant (TA) and low-voltage irregular (LVI) sleep (with scaling). The % proportion of epochs, with respect to the true state are shown in brackets. (d) Corresponding confusion matrix between clinician labels and the HMM (with scaling).

GMM, that does not account for state transitions. Both models also clearly outperform the unsupervised CLASS algorithm. While CLASS has previously shown to perform classification at term-age reasonably well, it does tend to perform less well when compared to younger ages [14].

The algorithm's lower kappa and sensitivity here suggests that QS is not completely detected. CLASS assumes that QS periods remain more discontinuous than non-QS. Although the TA periods of QS (of high discontinuity) is successfully classified, HVS periods (which are more continuous

Table 4. Results for four-state classification. Comparisons are made between GMMs and HMMs, with and without feature scaling.

Model	Mean kappa (SD)	Mean % acc. (SD)	Mean % sens. (SD)	Mean % spec (SD)	p_{boot}
HMM	0.62 (0.16)	86 (6.2)	72 (11)	91 (3.6)	—
GMM	0.55 (0.15)	82 (6.8)	65 (9.1)	88 (5.7)	0.25
HMM (no scaling)	0.56 (0.18)	84 (7.8)	62 (12)	90 (4.7)	0.32
GMM (no scaling)	0.51 (0.15)	81 (8.4)	59 (11)	87 (6.5)	0.19

acc: accuracy, sens: sensitivity, spec: specificity. SD: standard deviation. p_{boot} = proportion of bootstrapped recordings where model performs better than the HMM (with scaling).

in nature) are missed by the algorithm, making CLASS less reliable for classifying term-age recordings. The supervised nature of the GMM and HMM, on the other hand, allows both these QS states to be classified well. When also comparing to existing methods in literature, the hybrid evolutionary approach of Gerla *et al* reports a QS detection (versus AS) accuracy of 96% [28], while the multiscale entropy approach of De Wel *et al* presents a QS-non-QS classification sensitivity and specificity of 90% (based on the Receiver Operating Characteristics curve) [29]. Each of these approaches are performed with different features, training-testing procedures, and datasets, preventing a perfect comparison with this study. However, the high classification performance of these methods, together with the GMM and HMM of this study, indicates that there is a clear distinction between QS and non-QS states, with the strongly discontinuous, burst behaviours of QS easily distinguished from the relatively more continuous non-QS [12, 25].

It is in multi-state classification of ASI, HVS, TA and LVI, that we see a clearer improvement in HMM performance compared to the other models, with the introduction of personalized feature scaling resulting in a more consistent classification across the recordings. The strong distinction between QS and non-QS limits the problem of inter-recording variability in two-state classification, however, the subtler differences between states in four-state classification are more compromised by this variability and the overlap between different states across recordings reduces the discriminatory power of the features.

By introducing a transition probability matrix, we improve classifications further by accounting for the propensity of certain sleep state transitions to occur more often than others. For instance, HVS predominantly precedes TA if QS is sustained [12, 15], reflected by the higher transition probability from HVS to TA in figure 5(b), than from TA to HVS. In tandem with the very high probabilities of self-transitions, this accounts for the improved classification performance (compared to the GMM). However, even by accounting for transitions and correcting for some of the inter-recording variability, the similarities between the QS and non-QS state pairs make classification at this level still challenging, as term-age marks a period of transitionary sleep where certain preterm states begin to mature into states that resemble adult sleep [12, 15, 17].

The American Academy of Sleep Medicine recently renewed their criteria for the scoring of neonatal sleep (in infants zero to two months of age) [12], however, there are no strict rules for scoring sleep from EEG at this age and it remains predominantly based on expert clinical experience, often by a single rater [23]. The approach outlined in this study, for instance, is the standard procedure in use in the NICU, UZ Leuven, Belgium, but this will vary from place to place. The fact of the matter is that visual classification remains dubious. Testing on multiple datasets and including multiple, independent raters could further improve the quality of EEG labelling and classification accuracy. Recently, the NeoGUARD group coordinated by the NICU, UZ Leuven, has developed a method to improve the interrater scoring of seizure detection in newborn babies [60], and we are hopeful that this will pave the way for a similar approach to EEG sleep staging.

In this study, artefacts were not considered and were removed during pre-processing. However, in the NICU, these artefacts will often contaminate the recordings and prevent successful classification, especially when dealing with a transition-based approach such as the HMM. It would be advantageous to incorporate an automated method for artefact removal (as we previously introduced for CLASS), and including such methods is the next step towards a robust algorithm that is suitable for the clinical setting [14].

The small data size is a further limitation of this study, due to the time required to visually classify each recording, particularly for more than two sleep states. A larger dataset is likely to further improve classifier performance and generalizability, though this limitation underpins why there is a need for an automated classification method in the first place.

From a feature standpoint, synchrony measures were not considered in this study, and such features that are channel-specific (and therefore cerebral region-specific) may provide additional discriminatory features by which to better separate these sleep states. Features were also extracted across all EEG channels, however, there remains a clinical interest in performing sleep classification on a reduced set of channels, due to the small scalp sizes of the babies.

HMMs and GMMs form part of a family of generative, probabilistic models. Such models can be extended to perform semi-supervised learning, incorporating unlabelled epochs in the training procedure. Such an approach could reduce the uncertainty associated with visual labelling, and incorporate further, unlabelled datasets.

The ultimate aim is to extend these models to classify sleep across a wide range of preterm and term ages, opening the possibility of directly estimating the 'brain-age' of an EEG recording [61]. Existing studies have already shown that some of the features selected here, such as rEEG and MSE, show strong maturational trends with age [29, 39], allowing an additional dependency on age, as well as sleep state. By identifying the most likely age that fits the overall sleep structure of the EEG, we can form direct estimates of the developmental age of the baby's brain. Identifying discrepancies between this estimate and the known PMA, could provide an even earlier indication of delayed maturation [7–11].

By proving the applicability of these models for term-age EEG sleep staging in this study, we hope to pave the way for such future expansions, of which there is great potential.

5. Conclusion

This study presents a novel application of generative models, namely the GMM and HMM, for four-state sleep classification of term-age babies using only multichannel EEG. The HMM proves to be a successful approach for the classification of ASI, HVS, TA and LVI, with a mean kappa of 0.62 ($\pm$0.16). This is developed using a large set of features, inspired from multiple studies on adults and babies, to better utilize the full potential of the EEG. By accounting for sleep state transitions, overall classification performance improves, while incorporating a method of personalized feature scaling helps to correct for some of the inter-recording variability. Such an approach can optimize the quality and timing of neonatal care in the NICU and minimize unnecessary disturbances to the baby's sleep. More importantly, this could help quantify abnormalities in the development of preterm and term-born babies. By proving the classification ability of these models in term-age recordings, future work aims to extend these methods across a wider range of preterm and term ages.

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 Centre grant 098461/Z/12/Z (Sleep, Circadian Rhythms and Neuroscience Institute), the RCUK Digital Economy Programme grant EP/G036861/1 (Oxford Centre for Doctoral Training in Healthcare Innovation), IWT Leuven Belgium grant TBM 110697-NeoGuard, Bijzonder Onderzoeksfonds KU Leuven (BOF): The effect of perinatal stress on the later outcome in preterm babies grant C24/15/036, IMEC funds 2017, and ERC Advanced Grant: BIOTENSORS 339804.

Each author declares no conflict of interest, real or perceived.

The research materials supporting this publication can be accessed, upon reasonable request, by contacting the corresponding author.

Appendix. Additional feature information

A.1. Fourier and wavelet decompositions

Frequency-domain spectral measures such as band powers and mean frequencies, were obtained from the Fourier transform and Welch periodogram, calculated using a Hamming window with 50% overlap.

For the DWT, we chose a debauchies-4 (db4) wavelet as it has previously been shown to perform well in other EEG analyses [34, 35, 52]. Wavelet decomposition was performed using quadrature transforms to obtain the Detail coefficients cD1–cD5 and Approximate coefficient cA5, from which the decomposed EEG bands D1–D5 and A5 were reconstructed [24, 34].

A.2. Multiscale entropy (MSE) measures

MSE measures are obtained by calculating the sample entropy at increasingly coarser grains (scales) of the EEG [62]. As well as calculating the first 10 raw MSE values as features, we also calculate more complex MSE-based features defined by De Wel *et al*, based on the multiscale curve of MSE versus scale (for scales 1–20). From this, the area under multiscale curve, average slope of the curve (across scales 1–5), average slope of the curve (across scales 6–20), and maximum MSE value are extracted, as defined in [29]. As this previous study includes term-age EEG recordings, we use the same parameters of embedding dimension $m = 2$, and tolerance $r = 0.2 \times$ standard deviation.

A.3. LL suppression value

The method of LL, developed by Koolen *et al* is derived from the Katz Fractal Dimension [63]. LL is a metric sensitive to periods of discontinuity and bursts, and provides a robust measure of the level of suppression in the EEG, known as the suppression value (supp), as defined in [40].

A.4. Range EEG (rEEG) measures

The rEEG can be described as an approximation of the amplitude-integrated EEG (aEEG). It was developed by Navakatikyan *et al* and estimates the upper value of the EEG peak-to-peak amplitude. The difference between maximum and minimum values in adjacent 2 s EEG segments are linearly connected and up-sampled to 8 Hz frequency, providing a continuous signal sensitive to burst periods and suppression, as defined in [41]. Metrics of mean, median, SD, IQR, 5th percentile (lower margin (LM)) and 95th percentile (higher margin (HM)) are calculated as features to summarize the rEEG signal [39, 41].

ORCID iDs

Kirubin Pillay https://orcid.org/0000-0001-9269-2881

References

[1] Aarnoudse-Moens C, Weisglas-Kuperus N, Van Goudoever J and Oosterlaan J 2009 Meta-analysis of neurobehavioral outcomes in very preterm and/or very low birth weight children *Pediatrics* **124** 717–28

[2] Back S and Miller S 2014 Brain injury in premature neonates: a primary cerebral dysmaturation disorder? *Ann. Neurol.* **75** 469–86

[3] Larroque B *et al* 2008 Neurodevelopmental disabilities and special care of 5-year-old children born before 33 weeks of gestation (the EPIPAGE study): a longitudinal cohort study *Lancet* **371** 813–20

[4] de Kieviet J F, Zoetebier L, van Elburg R M, Vermeulen R J and Oosterlaan J 2012 Brain development of very preterm and very low-birthweight children in childhood and adolescence: a meta-analysis *Dev. Med. Child Neurol.* **54** 313–23

[5] Graven S and Browne J 2008 Sleep and brain development. The critical role of sleep in fetal and early neonatal brain development *Newborn Infant Nurs. Rev.* **8** 173–9

[6] Graven S 2006 Sleep and brain development *Clin. Perinatol.* **33** 693–706

[7] Scher M S, Steppe D A, Dahl R E, Asthana S and Guthrie R D 1992 Comparison of EEG sleep measures in healthy full-term and preterm infants at matched conceptional ages *Sleep* **15** 442–8

[8] Tokariev A, Videman M, Palva J M and Vanhatalo S 2015 Functional brain connectivity develops rapidly around term age and changes between vigilance states in the human newborn *Cereb. Cortex* **26** 4540–50

[9] Scher M S, Johnson M W, Ludington S M and Loparo K 2011 Physiologic brain dysmaturity in late preterm infants *Pediatr. Res.* **70** 524–8

[10] Biagioni E, Boldrini A, Giganti F, Guzzetta A, Salzarulo P and Cioni G 2005 Distribution of sleep and wakefulness EEG patterns in 24 h recordings of preterm and full-term newborns *Early Hum. Dev.* **81** 333–9

[11] dos Santos A Á, Khan R L, Rocha G and Nunes M L 2014 Behavior and EEG concordance of active and quiet sleep in preterm very low birth weight and full-term neonates at matched conceptional age *Early Hum. Dev.* **90** 507–10

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

[13] Mirmiran M, Maas Y G and Ariagno R L 2003 Development of fetal and neonatal sleep and circadian rhythms *Sleep Med. Rev.* **7** 321–34

[14] 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

[15] Husain A M 2005 Review of neonatal EEG *Am. J. Electroneurodiagn. Technol.* **45** 12–35

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

[17] 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

[18] 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 *Clin. Neurophysiol.* **40** 59–124

[19] Werth J, Atallah L, Andriessen P, Long X, Zwartkruis-pelgrim E and Aarts R M 2017 Unobtrusive sleep state measurements in preterm infants—a review *Sleep Med. Rev.* **32** 109–22

[20] Ludington-Hoe S M, Johnson M, Morgan K, Lewis T, Gutman J, Wilson P and Scher M 2006 Neurophysiologic Assessment of Neonatal Sleep Organization: preliminary results of a randomized, controlled trial of skin contact with preterm infants *Pediatrics* **117** e909–23

[21] Scher M S, Steppe D A, Banks D L, Guthrie R D and Sclabassi R J 1995 Maturational trends of EEG-sleep measures in the healthy preterm neonate *Pediatr. Neurol.* **12** 314–22

[22] Scher M S 2008 Ontogeny of EEG-sleep from neonatal through infancy periods *Sleep Med.* **9** 615–36

[23] 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

[24] Turnbull J, Loparo K, Johnson M and Scher M 2001 Automated detection of Tracé Alternant during sleep in healthy full-term neonates using discrete wavelet transform *Clin. Neurophysiol.* **112** 1893–900

[25] Dereymaeker A, Pillay K, Vervisch J, De Vos M, Van Huffel S, Jansen K and Naulaers G 2017 Review of sleep-EEG in preterm and term neonates *Early Hum. Dev.* **113** 87–103

[26] Shellhaas R, Burns J, Barks J and Chervin R 2014 Quantitative sleep stage analyses as a window to neonatal neurologic function *Neurology* **82** 390–5

[27] Hoppenbrouwers T, Hodgman J, Rybine D, Fabrikant G, Corwin M, Crowell D and Weese-Mayer D 2005 Sleep architecture in term and preterm infants beyond the neonatal period: the influence of gestational age, steroids, and ventilatory support *Sleep* **28** 1428–36

[28] Gerla V, Bursa M, Lhotska L, Paul K and Krajca V 2007 Newborn sleep stage classification using hybrid evolutionary approach *Int. J. Bioelectromagn.* **9** 25–6

[29] De Wel O, Lavanga M, Dorado A, 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

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

[31] Murphy K, Stevenson N J, Goulding R M, Lloyd R O, Korotchikova I, Livingstone V and Boylan G B 2014 Automated analysis of multi-channel EEG in preterm infants *Clin. Neurophysiol.* **126** 1692–702

[32] Carrozzi M, Accardo A and Bouquet F 2004 Analysis of sleep-stage characteristics in full-term newborns by means of spectral and fractal parameters *Sleep* **27** 1384–93

[33] Yetton B D, Niknazar M, Duggan K A, McDevitt E A, Whitehurst L N, Sattari N and Mednick S C 2016 Automatic detection of rapid eye movements (REMs): a machine learning approach *J. Neurosci. Methods* **259** 72–82

[34] Şen B, Peker M, Çavuşoğlu A and Çelebi F V 2014 A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms *J. Med. Syst.* **38** 18

[35] Greene B R, Faul S, Marnane W P, Lightbody G, Korotchikova I and Boylan G B 2008 A comparison of quantitative EEG features for neonatal seizure detection *Clin. Neurophysiol.* **119** 1248–61

[36] Temko A, Thomas E, Marnane W, Lightbody G and Boylan G 2011 EEG-based neonatal seizure detection with support vector machines *Clin. Neurophysiol.* **122** 464–73

[37] Li S, Zhou W, Yuan Q, Geng S and Cai D 2013 Feature extraction and recognition of ictal EEG using EMD and SVM *Comput. Biol. Med.* **43** 807–16

[38] 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

[39] 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

[40] Dereymaeker A, Koolen N, Jansen K, Vervisch J, Ortibus E, De Vos M, Van Huffel S and Naulaers G 2016 The suppression curve as a quantitative approach for measuring

brain maturation in preterm infants *Clin. Neurophysiol.* **127** 2760–5
[41] Navakatikyan M A, O'Reilly D and Van Marter L J 2016 Automatic measurement of interburst interval in premature neonates using range EEG *Clin. Neurophysiol.* **127** 1233–46
[42] Krajča V, Petránek S, Mohylová J, Paul K, Gerla V and Lhotská L 2007 Neonatal EEG Sleep Stages Modelling by Temporal Profiles *Computer Aided Systems Theory—EUROCAST 2007* (Berlin: Springer) pp 195–201
[43] Löfhede J, Löfgren N, Thordstein M, Flisberg A, Kjellmer I and Lindecrantz K 2008 Classification of burst and suppression in the neonatal electroencephalogram *J. Neural Eng.* **5** 402–10
[44] Huang R S, Tsai L L and Kuo C J 2001 Selection of valid and reliable EEG features for predicting auditory and visual alertness levels *Proc. Natl Sci. Counc.* **25** 17–25
[45] Li C 2013 Complexity analysis of physiological time series with applications to neonatal sleep electroencephalogram signals *PhD Thesis* Case Western Reserve University, Cleveland, OH, USA
[46] 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
[47] Li Y, Yingle F, Gu L and Qinye T 2009 Sleep stage classification based on EEG Hilbert-Huang transform *2009 4th IEEE Conf. on Industrial Electronics and Applications* (IEEE) pp 3676–81
[48] Fraiwan L, Lweesy K, Khasawneh N, Wenz H and Dickhaus H 2012 Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier *Comput. Methods Programs Biomed.* **108** 10–9
[49] Kaffashi F, Scher M S, Ludington-Hoe S M and Loparo K A 2013 An analysis of the kangaroo care intervention using neonatal EEG complexity: a preliminary study *Clin. Neurophysiol.* **124** 238–46
[50] Scher M S, Dokianakis S G, Sun M, Steppe D A, Guthrie R D and Sclabassi R J 1996 Computer classification of sleep in preterm and full-term neonates at similar postconceptional term ages *Sleep* **19** 18–25
[51] Estévez P A, Held C M, Holzmann C A, Perez C A, Pérez J P, Heiss J, Garrido M and Peirano P 2002 Polysomnographic pattern recognition for automated classification of sleep-waking states in infants *Med. Biol. Eng. Comput.* **40** 105–13
[52] Temko A, Stevenson N, Marnane W, Boylan G and Lightbody G 2012 Inclusion of temporal priors for automated neonatal EEG classification *J. Neural Eng.* **9** 46002
[53] Wang S, Gwizdka J and Chaovalitwongse W A 2016 Using wireless EEG signals to assess memory workload in the n-back task *IEEE Trans. Hum. Mach. Syst.* **46** 424–35
[54] Tsanas A, Little M A and Mcsharry P E 2013 A methodology for the analysis of medical data *Handbook of Systems and Complexity in Health* (New York: Springer) ch 7, pp 113–25
[55] McLachlin G and Peel D 2000 *Finite Mixture Models* (New York: Wiley)
[56] Vaseghi S 2008 *Advanced Digital Signal Processing and Noise Reduction* (New York: Wiley)
[57] Mandrekar J N 2011 Measures of interrater agreement *J. Thorac. Oncol.* **6** 6–7
[58] Gwet K 2012 *Handbook of Inter-Rater Reliability: the Definitive Guide to Measuring the Extent of Agreement among Multiple Raters* (Gaithersburg, MD, USA: Advanced Analytics, LLC)
[59] Efron B 1979 Bootstrap methods: another look at the jackknife *Ann. Stat.* **7** 1–26
[60] Dereymaeker A *et al* 2017 Interrater agreement in visual scoring of neonatal seizures based on majority voting on a web-based system: the Neoguard EEG database *Clin. Neurophysiol.* **128** 1737–45
[61] O'Toole J, Boylan G, Vanhatalo S and Stevenson N 2016 Estimating functional brain maturity in very and extremely preterm neonates using automated analysis of the electroencephalogram *Clin. Neurophysiol.* **127** 2910–8
[62] Costa M, Goldberger A L and Peng C-K 2005 Multiscale entropy analysis of biological signals *Phys. Rev.* E **71** 21906
[63] Katz M J 1988 Fractals and the analysis of waveforms *Comput. Biol. Med.* **18** 145–56