PAPER

Automated preterm infant sleep staging using capacitive electrocardiography

IOP Publishing Physiol. Meas. 40 (2019) 055003 (14pp) https://doi.org/10.1088/1361-6579/ab1224

Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

PAPER

Automated preterm infant sleep staging using capacitive electrocardiography

RECEIVED
11 December 2018

REVISED
14 March 2019

ACCEPTED FOR PUBLICATION
21 March 2019

PUBLISHED
4 June 2019

Jan Werth[1], Aline Serteyn[1,2], Peter Andriessen[3], Ronald M Aarts[1,2] and Xi Long[1,2]

[1] Department of Electrical Engineering, University of Technology Eindhoven, Eindhoven, The Netherlands
[2] Philips Research, Eindhoven, The Netherlands
[3] Paediatric Department, Máxima Medical Center, Veldhoven, The Netherlands

E-mail: werth.jan@gmail.com **and** xi.long@philips.com

Keywords: automated sleep classification, capacitive ECG, classic ECG, machine learning, preterm infants, unobtrusive

Abstract

Objective: To date, mainly obtrusive methods (e.g. adhesive electrodes in electroencephalography or electrocardiography) have been necessary to determine the preterm infant sleep states. As any obtrusive measure should be avoided in preterm infants because of their immature skin development, we investigated the possibility of automated sleep staging using electrocardiograph signals from non-adhesive capacitive electrocardiography. *Approach*: Capacitive electrocardiography data from eight different patients with a mean gestational age of 30 $\pm$ 2.5 weeks are compared to manually annotated reference signals from classic adhesive electrodes. The sleep annotations were performed by two trained observers based on behavioral observations. *Main results*: Based on these annotations, classification performance of the preterm infant active and quiet sleep states, based on capacitive electrocardiography signals, showed a kappa value of 0.56 $\pm$ 0.20. Adding wake and caretaking into the classification, a performance of kappa 0.44 $\pm$ 0.21 was achieved. In-between sleep state performance showed a classification performance of kappa 0.36 $\pm$ 0.12. Lastly, a performance for all sleep states of kappa 0.35 $\pm$ 0.17 was attained. *Significance*: Capacitive electrocardiography signals can be utilized to classify the central preterm infant sleep states, active and quiet sleep. With further research based on our results, automated classification of sleep states can become an essential instrument in future intensive neonatal care for continuous brain maturation monitoring. In particular, being able to use capacitive electrocardiography for continuous monitoring is a significant contributor to reducing disruption and harm for this extremely fragile patient group.

List of abbreviations

ADASYN	Adaptive synthetic sampling
AS	Active sleep
CPAP	Continuous positive airway pressure
CTW	Caretaking and wake
DAQ	Data acquisition (system)
DT	Decision tree
ECG	Electrocardiogram
EEG	Electroencephalogram
ERF	Extra tree random forest
GA	Gestational age
GB	Gradient boosting
HRV	Heart rate variability
IS	Intermediate sleep
LL	Line length
LOOCV	Leave one out cross validation
LZ	Lempel-Ziv complexity measure

NICU	Neonatal inensive care unite
NN	Normal-to-normal beat
PMA	Postmenstrual age
QS	Quiet sleep
QSE	Quadratic sample entropy
RBF	Radial basis function
RF	Random forest
SDLL	Standard derivation of line length
SE	Sample entropy
SEAUC	Sample entropy area under the curve
SER	SE over a range of r values
SMOTE	Synthetic minority over-sampling technique

1. Introduction

Preterm infant sleep is strongly connected to brain maturation. A more in-depth explanation of the importance of sleep on brain maturation and the classification of sleep in preterm infants is reviewed elsewhere (Werth *et al* 2017a). In short, sleep in newborns can be separated into three sleep states, active sleep (AS), quiet sleep (QS) and intermediate sleep (IS) and wake and/or caretaking. AS is known to activate neuronal activities and is essential in synaptogenesis of the neural interconnections. AS is the most dominant state with around 70% of the total sleep time in the first weeks after birth. During QS, neuronal activity is also seen but less than in AS. QS is mostly described as a resting or reenergizing state. Also, developmental errors are corrected during QS using the heightened brain plasticity of the preterm infants to reorganize the brain structure. In the preterm brain, the time spent in AS and QS fluctuates quickly and may be difficult to separate during the first weeks after birth. Many episodes cannot be explicitly allocated to one specific state and therefore IS is more prominent at that early state in preterm infants. The distribution of the sleep states can be a biomarker of brain maturation.

In clinical practice, sleep staging is mainly based on manually annotated electroencephalogram (EEG) and/or behavioral analysis. These practices are time-consuming and not continuous. Therefore, several research groups work on automated sleep staging algorithms for preterm infants (Scher *et al* 2005, Gerla *et al* 2007, Isler *et al* 2016, Dereymaeker *et al* 2017, Koolen *et al* 2017, Werth *et al* 2017b). To date, the research is mainly focused on EEG and electrocardiography (ECG) analysis. For both signal types, standard adhesive electrode settings are used. As mentioned in previous publications, all electrodes in contact with the fragile preterm infant skin are considered obtrusive (Gruetzmann *et al* 2007, Werth *et al* 2017a). The epidermis of a preterm infant under 32 weeks' of gestation is only two to three layers thick with almost no protective outer skin. Disrupting the immature epidermis results in an increased risk of infection, healing, and scarring (Atallah *et al* 2014). In a recent paper, we reviewed different unobtrusive methods for the use of preterm infant sleep staging (Werth *et al* 2017a).

In this publication, the focus is on the use of ECG measurements from capacitive ECG (cECG) electrodes to determine different sleep states in preterm infants automatically. It will be determined which R-peak detection method is superior for cECG analysis. Secondly, it will be investigated how well the previously used features perform in comparison to newly implemented features for AS and QS separation. Then, it will be investigated how multi-class classification performs comparing the use of the ECG and cECG signals. Finally, the difference in the successfully used features for the ECG and cECG classification will be investigated.

2. Methods

2.1. Population

In this retrospective study, eight stable preterm infants born with a mean gestational age (GA) of 30 $\pm$ 2.5 weeks were analyzed. They were studied at a mean postmenstrual age (PMA) of 32 $\pm$ 2.6 weeks. The patients had a mean birth weight of 1652 $\pm$ 565 g. They were admitted to the neonatal intensive care unit (NICU) of the neonatal department at the Máxima Medical Center Veldhoven, The Netherlands. Ethical approval was given by the medical ethical committee of the hospital; written consent was given by the patient's parents.

2.2. Annotations

The data was annotated by two trained observers based on 30 s epochs adhering the Prechtl system (Prechtl 1974). The observers used a reference ECG time series and video information for annotation. Their adapted annotation style was tested in another trial and proven to be on par with gold standard full PSG annotations (Otte and Long 2019). They annotated the following states: AS, QS, IS, wake, caretaking, and unknown (unable to annotate). The total duration of annotated data was 40 h (4850 30 s epochs) with a mean duration per patient of 5.2 $\pm$ 1.3 h (630 $\pm$ 157 30 s epochs). The overall distribution of state was: AS: 62.7%, QS: 8.2%, IS: 13.7%, wake:

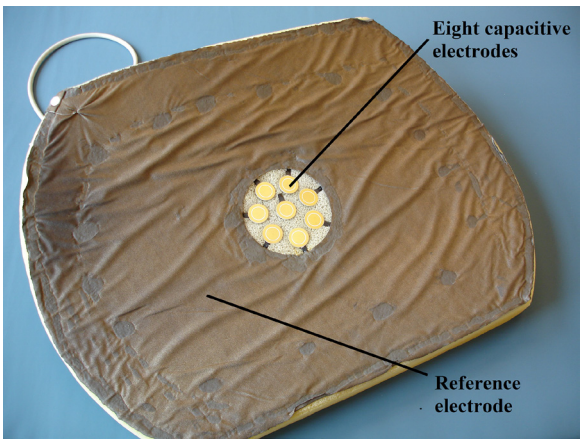

Figure 1. Incubator mattress with eight capacitive sensors and a conductive cover material acting as the reference electrode. The mattress would be protected with a polyurethane cover to withstand fluids (e.g. urine) and for easy cleaning and disinfection.

2%, caretaking: 11.4% and unknown: 1.9%. The mean interrater reliability was high with kappa of 0.70. The highest agreement was found in AS, QS and caretaking with a kappa of 0.75, 0.74, and 0.71. Caretaking was not perfectly agreed on mainly due to some difficulties to determine the exact beginning and end of the caretaking procedure. While the differences were found mainly in-between states with different start and duration of the transition state resulting in a kappa of 0.59. Wake was agreed on with a kappa of 0.69. After clarifying differences in concordance with a third trained annotator, the observers reached consent. The concordant annotations were used for further analysis. The Cohen's kappa statistic defines a score below 0.41 as poor, over 0.41 as moderate, over 0.61 as substantial, and over 0.81 as almost perfect (McHugh 2012).

As preterm infants are mostly awake during caretaking periods, generating very similar signal structures, the labels caretaking and wake were merged under the label caretaking + wake (CTW).

2.3. Data recordings

For each infant, ECG, cECG, and videos were recorded in three to five sessions within one week. The reference ECG was recorded with three standard leads via a Philips Monitor (Intellivue Mx800, Germany) at a sampling frequency of 500 Hz. The cECG sensors where connected to a data acquisition system (DAQ) system and the cECG data were recorded at 8 kHz, then downsampled to 500 Hz. The videos were recorded with a standard, medium resolution, grayscale camera. The camera was mounted either for facial view or total body view. All signals were time synchronized to the video recordings.

2.4. Capacitive ECG

To obtain the cECG, a special incubator mattress with eight capacitive sensors was used. The electrodes were connected to the DAQ. The top layer of the mattress itself was made out of conductive material to function as a reference to the eight capacitive electrodes (figure 1). The mattress was covered with a polyurethane cover to withstand fluids (e.g. urine) and for easy cleaning and disinfection. The polyurethane layer was then covered with a cotton bed sheet for comfort and to reduce the triboelectric effect appearing on both electrodes types (Gordon 1975). The triboelectric effect creates an electric charge when friction is applied between two different materials, e.g. due to movement of the neonate, and is not the single of interest but rather an artifact. The baby should be positioned as to cover both, at least a part of the sensors array and a part of the reference electrode. The position (supine, side, prone) does not influence the heart rate determination but changes the shape of the ECG signals. To provide further comfort, the neonate was placed on the mattress in a cotton snuggle.

A vectorcardiogram and its projection on the Einthoven leads (Becker 2006), were constructed from the raw data after a channel selection process. Before the channel selection, neutralization was used on the analog signal eliminating the input impedance of the sensor amplifier to reduce motion induced noise (Veen *et al* 2011). The channel selection process first ranked the channels by its coupling factor between sensors and body. The coupling factor was determined by injecting a low-current 1 kHz, whose signal amplitude decay is proportionate to the coupling, into the electrodes (Serteyn *et al* 2015). Secondly, bad channels, i.e. those with a variance higher than a certain threshold, were eliminated. For details about the channel selection, we refer to Atallah *et al* (2014). The ECG was then reconstructed from at least three selected channels (Vullings *et al* 2010).

The signals from the selected channels were first down-sampled, and a bandpass filter of 3 to 35 Hz was applied, thus removing all frequency components outside the (dominant) ECG band, e.g. the well-known

50/60 Hz common-mode interference was rejected. To subtract the common mode, which is significantly affected by people walking by, the signals were averaged and the average signal subtracted from each channel signal.

2.5. *R* peak detection

To analyze the heart rate variability (HRV) from the cECG, R-peak detection and normal-to-normal beat (NN) interval determination are very important. Slight variations in peak detection would introduce false sleep staging as the difference between the sleep states is only minimal. In this paper, we compare two R peak detection methods to determine which yield better results for the presented data. One method is from Wijshoff *et al* (2017) and a second from Rooijakkers *et al* (2012). The algorithm from Wijshoff *et al* was not initially intended for preterm infants but was confirmed to be working well in this patient group (Werth *et al* 2017b). Rooijakkers' method was created and confirmed for fetal monitoring.

Wijshoff *et al* calculated the first derivative of the ECG signal to search for steepest ascent and descent of the QR and RS slopes. A variable threshold was applied to detect the peaks in the QRS complex. They then used a sub-peak detection to verify the peak position at the real max by interpolating around the found peaks. The sub-peak detection assured that there is no shift from the real peak due to off sampling. Rooijakkers *et al* band passed the ECG signal locally with the use of time discrete continuous wavelet transform with the peak wavelet frequency centered in the 10–25 Hz frequency band. They then segmented the ECG signal to obtain one QRS complex per segment. Those segments are run through a variable threshold to find the R peaks within a set time window which is based on the previously found R peak.

Cross-correlation was determined between the HRV signal created from the ground truth ECG and the cECG signal for each R peak detection method. The features for sleep state classification were created for both methods, and the classification performance was compared using the two different methods.

2.6. Features

To be able to separate the different states, in total 34 ECG and HRV features in the time, frequency and non-linear domain were determined. The features were calculated on the base of 300 s windows centered on 30 s epochs. The features are calculated for each 30 s epoch and averaged over the corresponding 300 s window. An overview of all used features can be found in table 1.

For the HRV, the needed ECG R-peaks are fundamentally non-equidistant in time. To avoid resampling the RR signal, and thereby introducing extra parameters, the Lomb–Scargle algorithm was applied to generate the frequency spectrum (Ruf 1999).

As a base, a set of existing HRV features which were already used in our previous publication was implemented (Werth *et al* 2017b). These are linear HRV features focusing mainly on direct representation of changes in the para-/sympathetic nervous system expressed in cardiorespiratory changes. To accommodate the increased cardiorespiratory rates in preterm infants, two preterm infant-specific features were added in the high-frequency domain, sHF, and uHF. The frequency band was extended based on adult sleep staging to 0.45–0.7 Hz for sHF and 0.7–1.5 Hz for uHF, and then the spectrum power in these two bands was computed.

To investigate the movement induced noise (e.g. body movement artifacts) of the capacitive electrodes, features calculated directly from the ECG signals (ECG and cECG) were introduced. Movement artifacts can be a direct and indirect indication for certain sleep states. Beats per epoch (BpE) counts the R peaks in an interval of 300 s. Line length (LL) and mean LL (aLL) calculates and averages the length of the ECG time series signal over a window of 300 s. Also, the standard derivation is calculated over LL (SDLL) and aLL (SDaLL) in 300 s windows. The LL is calculated with linear piecewise approximation using numerical integration for each segment of 30 s to calculate the arc length which is then summed up over 300 s epochs to gain the cumulative chordal distance.

Kommers *et al* (2017) designed two new features specifically for preterm infants to capture regulatory changes during kangaroo care: the percentage of HR decelerations (pDec) and the magnitude of HR deceleration (SDDec). The two features indicate how many HR decelerations occur and the extent of those decelerations. They showed that pDEC and SDDec were strongly affected by kangaroo care supporting that HR decelerations are a consequence of the autonomic nervous system response. Thereby, these features are expected to be influenced by sleep state changes.

Lucchini *et al* (2016) suggested that non-linear sample entropy (SE) and quadratic sample entropy (QSE) describe the preterm infant autonomic response. As the autonomic response is directly linked to the sleep states, those non-linear features were incorporated. In the literature, it is described that the standard value for the tolerance parameter r of 20% times the standard derivation is not accurate enough anymore, especially in preterm infants where instabilities are the norm (Lake 2011, Yentes *et al* 2013). Therefore, the SE was calculated on an adaptive tolerance parameter r. The embedded dimension m was fixed to 2 following (Yentes *et al* 2013, Lucchini *et al* 2016). The r value per epoch was determined by calculating the SE over a range of r from 0.05 to 0.3 times the standard derivation (SER) for 300 s epochs, which also includes the standard value for r. The calculation to find the optimal r value was done as described in the following points:

Table 1. Overview of the used ECG and HRV features for classification.

NR	Feature [unit]	Description
0	BpE [count]	Beats per epoch
1,2	LL, aLL [mV]	Line length/mean line length
3–6	NNx [count]	The number of pairs of successive R–R intervals that differ by more than 10, 20, 30 or 50 ms of a defined window length.
7–10	pNNx [%]	The proportion of NNx divided by total number of R–R intervals of a defined window length.
11	RMSSD [ms]	Root mean square of successive differences between adjacent R–R intervals of a defined window length.
12	SDALL [mV]	Standard derivation of averaged line length
13	SDANN [ms]	Standard deviation of averaged NN intervals
14	SDLL [ms]	Standard derivation of line length
15	SDNN [ms]	The standard deviation of normal to normal R–R intervals of a defined window length
16	HF [ms^2]	The power of the high frequency band between 0.15–0.4 Hz of a defined window size.
17	HFnorm [%]	HF power in normalized units HF/(total power-VLF) × 100
18	LF [ms2]	The power of the low frequency band between 0.04–0.15 Hz of a defined window size
19	LFnorm [%]	LF power in normalized units LF/(total power-VLF) × 100
20	LF/HF [n.u.]	Ratio LF/HF
21	sHF [ms^2]	The power of the high frequency band between 0.4–0.7 Hz
22	sHFnorm [%]	sHF power in normalized units sHF/(total power-VLF) × 100
23	TotPow [ms^2]	Total power or variance of NN intervals of a defined window size
24	uHF [ms^2]	The power of the high frequency band between 0.7–1.5 Hz
25	uHFnorm [%]	uHF power in normalized units uHF/(total power-VLF) × 100
26	VLF [ms^2]	The power of the very low frequency band between 0.003–0.04 Hz of a defined window size
27,28	SEN, QSE [n.u.]	Sample entropy/quadratic sample entropy
29	SEAUC [n.u.]	Sample entropy area under the curve
30	pDEC [%]	The percentage of HR decelerations
31	SDDec [ms]	Magnitude of HR deceleration
32,33	LZNN [n.u.], LZECG [n.u.]	Lempel-Ziv complexity measure on HRV and ECG

- Calculate SE over a range of *r* values (SER).
- Generalize the SER curve by low pass filtering (e.g. moving average).
- Find drop off/turning point of the SER curve where the amount of matches increases and entropy decreases.
- Find the minimum value of the SER curve.
- Calculate the mean SE value between the SER curve turning point and the minimum value.
- Find the *r* with the closest entropy value to the calculated mean *r*-value in the original SER curve.

In addition to SE and QSE, another fluctuation measure was calculated: the Lempel-Ziv complexity measure (LZ) for the ECG (or cECG) and corresponding HRV signal. With LZ, the different signal structure is measured using the ECG (or cECG) and HRV time series as analytic signals. The LZ is more prone to signal length than SE and QSE, but when used on a fixed time window the effect will not have any influence on the calculated LZ values. Another novel feature is the sample entropy area under the curve (SEAUC). Here the previous calculated SER curve is extended to a range of *r* of 0.1 to 20 creating a longer tail. Then the curve is integrated to show the difference between the curve shapes for different sleep states. With a higher entropy in the signal, matches are found later with a higher tolerance *r*, shifting the drop off/turning point.

2.7. Preprocessing

Before feeding the features to the learning routine, preprocessing steps were performed. The data were first normalized per recorded session using the Python scikit-learn standard-scaler and Min-Max-scaler for comparison (Pedregosa *et al* 2011). Selected features were averaged per session with a moving average using different window lengths. As the data was highly unbalanced, the synthetic minority over-sampling technique (SMOTE) (Chawla *et al* 2002) and adaptive synthetic sampling approach (ADASYN) (He *et al* 2008) were applied to increase the number of data points for more stable and generalized learning. To be able to distinguish

non-linear and inter-feature correlations and to increase linear and non-linear separability of the sleep states, feature transformation was applied to elevate the feature space to a higher dimension. A second order polynomial feature transformation and radial basis function kernel (RBF) were used with gamma grid search between 0.001 and 10 to transform the selected features. The polynomial function for feature F_x results in a new feature set: F_x, $F_{x*}F_y$, F_x^2. The additional created features F_y, F_y^2 were removed as only the transformation of the selected features F_x are of interest. The not intended feature transformation F_y, F_y^2 could lead to overfitting, and the redundant information could decrease the classification performance.

Further, the input parameters were averaged with a moving window. The size of the moving window was chosen between 0 and 50 30 s epochs to incorporate short and long-term averaging factors.

The data was separated into training, validation and test sets splitting by patients to be used later in a leave one patient out cross-validation (LOOCV). As the data set is small, a classic data splitting based on a fraction into training, validation and test set could not be performed as it would have introduced bias and lead to overfitting on the classifier side. To avoid bias and overfitting, it was chosen to split the data between patients to assure that the validation and especially testing set represents unseen information. In the next step, a classifier is empirically chosen for each class-set. The classifier is fed with its parameters and preprocessed data. The performance is then validated with LOOCV, and the classifier parameters and preprocessing are adapted to optimize the classification performance.

2.8. Selection path strategy

As only a small group of patients was available with an unbalanced state distribution, a selection path model for each state was implemented. The goal is to increase the multiclass classification performance by separating the classification in smaller sub-classification problems which can be optimized with the use of a wrapper including data preprocessing, different classifiers, feature selection, and parameter tuning. Also, by creating those intermediate classifications, it can be determined which factors influence the classification of a particular state. The final target is a full class classification.

Using intermediate steps until full class classification, the classification was separated into smaller groups of classes to identify the classification performance for individual state groups: AS-QS, AS-QS-IS, AS-QS-CTW. The classification of each state-group focuses on a sub-state-group composed of one minority state (QS, IS, CTW) and the majority state AS, e.g. AS-QS out of AS-QS-IS. Within the wrapper, the chosen classifier, feature selection, and parameters were individually optimized for each of the sub-state-groups. Each sub-state-group classification results in a prediction. Later, the wrapper merges the sub-state-predictions under a ruleset to a final prediction per state-group as wrapper output (figure 2).

The ruleset is such constructed that if AS, as the majority state, is below a minimum probability threshold, the state with the highest probability among the minority states is chosen. This ruleset is performed for each class-set probability output which is optimized for a specific minority state. The minimum probability threshold is determined via a grid search. The class-set optimization uses the F1 score on the validation sets. The implemented probability cut off for AS is used to limit the influence of the majority class AS. Finally, the class-set optimized predictions are merged. Upper and lower probability ruling thresholds decide which sleep state is for the final class prediction per epoch. The whole process is assessed by evaluating the final sleep state identifications from the test set against the annotations using the kappa score (κ) (McHugh 2012).

2.9. Classification

Five different classifiers from the Python scikit-learn library (Pedregosa *et al* 2011) are used for the wrapper to choose, which are decision tree (DT), RBF kernel support vector machine (SVM), random forest (RF), extra tree random forest (ERF) and gradient boosting (GB) (Duda *et al* 2000, Natekin and Knoll 2013).

The classifiers' input parameters are specific for different kind of classifiers. The adapted input parameters for the tree learners DT, RF, ERF were the number of trees, measure of branch splitting quality, depth of the tree, the minimum amount of samples representing a leaf node and the minimal number of samples allowing for a split. Also, the GB tree can be run with logistic regression or AdaBoosting loss function. Another parameter is the learning rate (also shrinkage or eta) which corrects for prediction errors from the existing trees by weighting each tree contribution. The SVM needs the misclassification penalty parameter C and the hyperparameter *gamma* which is determining the influence of the support vector on the class decision. Both parameters were found with a pre-grid search including all features to find the overall optimal values. As a starting point, the values 3 for C and 3.8 for *gamma* were used as good balance between speed and accuracy without losing generalization properties.

2.10. Feature selection and parameter determination

To choose the right feature set and classifier parameters for each separate sleep state classification, an adapted greedy-bi-directional (backward and forward) search was implemented. In the first step, all features are fed into the system without any dimension expansion (by feature transformation). After classification on the validation

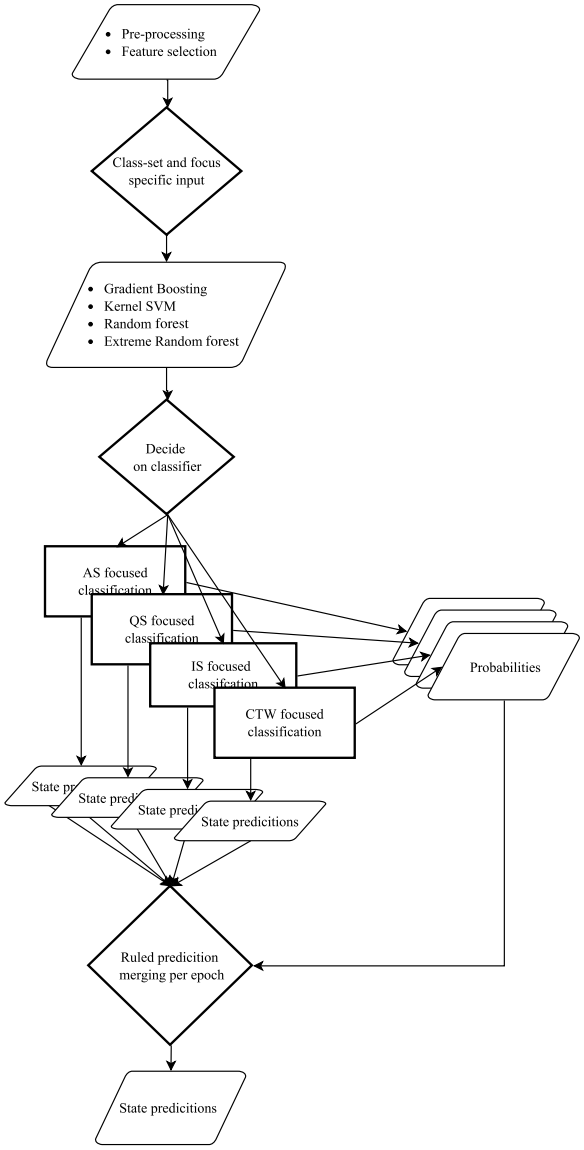

Figure 2. Overview of the multi-classifier approach. In the first step, parameters for each class are selected separately which are then fed to a chosen classifier. The classifier predictions for each state are merged into a single prediction using the state probability for each epoch. The merging uses a ruleset using the probabilities per class-set to decide on the final state prediction. The outcome are joint state/label predictions.

set, the feature importance is determined in the first validation run, and the best features are boosted with dimension expansion and/or the worst features can be removed. The classification performance is determined, and the feature transformation and/or the feature removal extended or reversed. The process is repeated until the performance on the validation set does not increase any further in the third decimal place. As the combinations of different parameters, classifiers, and chosen/boosted features are extraordinaire, this selection was not executed via an automated exhaustive grid search but partially manually with experimental alteration of feature selection and best first greedy search for parameter optimization. An overview of the adapted parameters can be found in table 3. A simplified overview of the whole process can be seen in figure 3.

3. Results

3.1. R peak coverage comparison for ECG and cECG

The R-peak intervals for the normalized ECG and cECG signals were determined per session, and the cross-correlation was calculated to see how well both signals align with their reference counterpart. The overall mean correlation per patient using the R-peak detector by Wijshoff *et al* is 0.63 ± 0.04. The method of Rooijakkers *et al* shows a slightly increased correlation of 0.692 ± 0.22 between the cECG and ECG R peaks. The discrepancy between the R-peaks detected from cECG, and the ECG mainly comes from the different impact of noise on the signal. Noise removal treatment did not enable reliable R peak detection and was therefore left out to avoid the

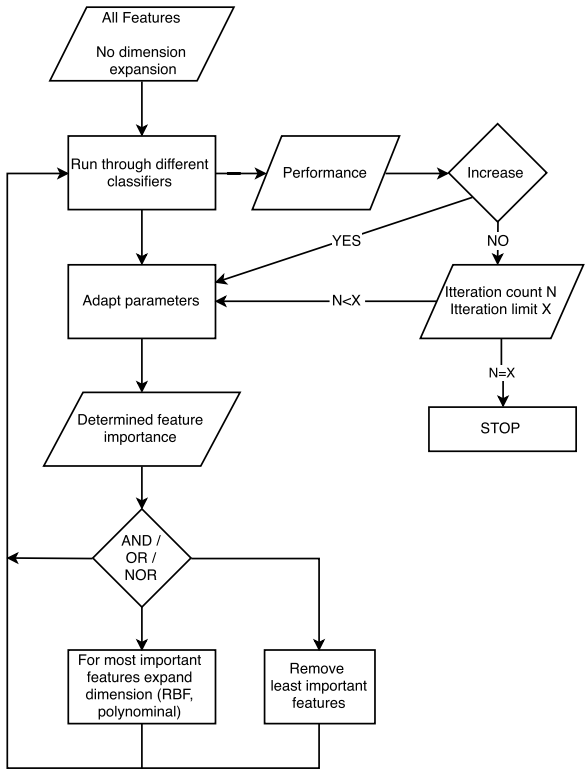

Figure 3. The flow diagram shows how the Features are selected. At start all Features are used without dimension expansion like *n* order polynomial or the Radian Bases function. After classification with a chosen classifier, the most important featured are transformed to increase their class distinction. The least important features can be dropped. A best first greedy approach was used to adapt the parameters with the chosen features. The parameters can be found in table 3. This is continued until the performance does not increase anymore.

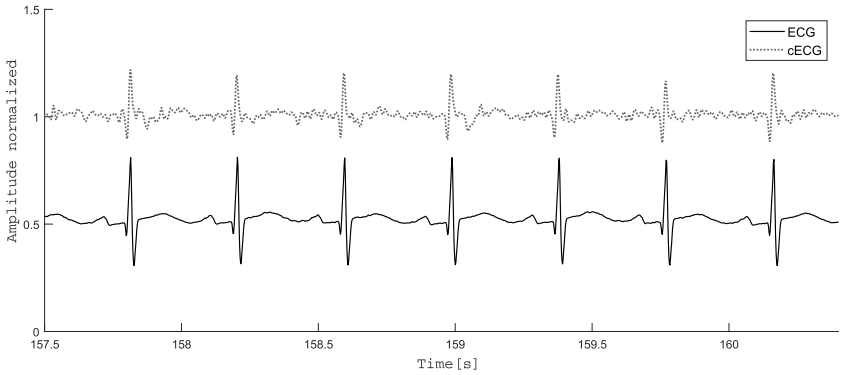

Figure 4. ECG and capacitive ECG (cECG) simultaneously recorded. The ECG signal is more distinct but the cECG shows good potential for R-peak detection. For a more discriminable display of signals the cECG was plotted with an amplitude offset.

creation of false signal episodes regarding R peak detection. Episodes, where the preterm infant was lying still and covering the electrodes, generate good ECG signals for peak detection (figure 4).

For other episodes nevertheless, the signals were corrupted with noise similar to the ECG in composition but without actual R-peaks (figure 5), creating partly false R-peak detections. Nevertheless, the different R-peak methods do not show any significant impact on sleep staging performance. To create consistency with our previous publication (Werth *et al* 2017b) the R-peak method of Wijshoff *et al* was used in this work.

3.2. Feature importance

The 34 features have different importance regarding the different state classification and for the used signal type of ECG or cECG. In figures 6 and 7 the feature importance for each sub-classification used for classification including all states is displayed. The features are displayed before feature transformation to recognize which main feature types are of key importance. To identify which feature types are important for ECG and which for cECG, the overall important features per signal are listed below. Overall important features are features exceeding for at

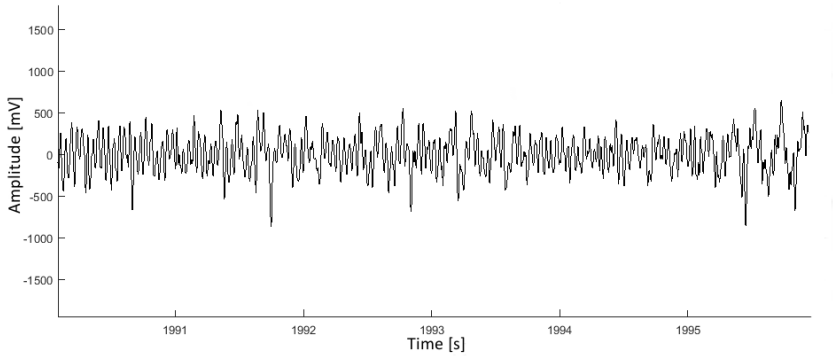

Figure 5. This image shows a cECG signal which can easily be mistaken by the R-peak detector for a signal with QRS complexes modulated with noise. The R-peak detection is not working properly on such epochs resulting in decreased correlation between ECG and cECG HRV signals.

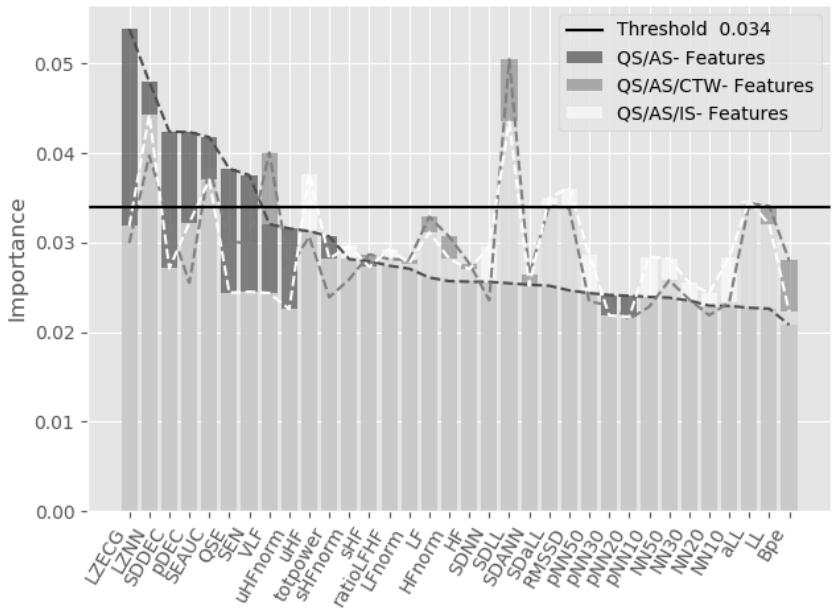

Figure 6. Feature importance for each class-set sorted after AS-QS most important features using the cECG signal. This figure represents the feature importance for classifying between AS, QS, CT/W and IS (state-group). Each of this feature sets was used to classify specifically focused on one class-set to later merge the predictions. The threshold determines which features are deemed as important. For the overall importance per state-group, features that exceeds the threshold for at least two sets are chosen.

least two subsets the threshold of 3.4% from the total distributed importance which sums up to a total of 1. The threshold was found by a sweep analysis resulting in the highest performance based on the ECG features.

These features are listed in table 2 per used state-groups and signal type. Here, only the last state-group AS-QS-CTW-IS corresponds to the displayed figures 6 and 7.

3.3. Parameter selection

The used parameter values varied for each task. Depending on the state combination, bi- or multi-state classification, and the used dataset, the performance changed based on the used parameter combination. The used parameter ranges can be found in table 3. Also, feature groups were left out in the averaging step or polynomial transformation. Also, not all features were always used (see table 4).

The parameters are presented in ranges as presenting all combinations in detail would exceed the purpose of this publication and would not benefit the reader as the specific parameter values are tied to the here used datasets.

3.4. Capacitive versus classic three lead ECG

To identify the sleep states in a genuinely unobtrusive manner, a cECG system was used to capture the ECG and HRV. Looking at AS and QS only, the analysis resulted in a κ of 0.42 ± 0.26; 0.49 (mean $\kappa \pm$ std; cumulative pooled κ) for the ECG and a κ of 0.56 ± 0.20; 0.51 for the cECG. The performance changed when more sleep

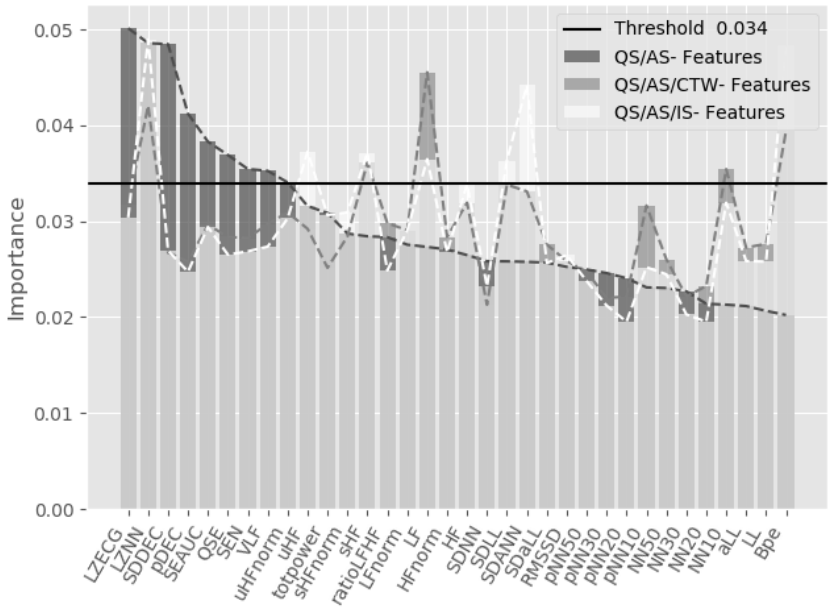

Figure 7. Feature importance for each class-set sorted after AS-QS most important features using the classic ECG signals. Framework same as in figure 6.

Table 2. Overview of most relevant features for classification per state-group.

State-group	Signal	Features
AS-QS	ECG	LF, SDANN, LZNN
	cECG	LL, aLL, SDANN, LZECG, SDLL, SEAUC, LZNN
AS-QS-CTW	ECG	BpE, LZNN, SDANN, HF, sHF, LF, SDLL, SDNN
	cECG	SDLL, SDaLL, LL, ZECG, aLL
AS-QS-IS	ECG	SDANN, BpE, LF, totpower, uHF, uHFnorm
	cECG	SDLL, LZECG, LL, SDaLL, aLL, LZNN, SEAUC
AS-QS-CTW-IS	ECG	LZNN, BpE, LF, sHF
	cECG	SDLL, LZNN, SDaLL, aLL, SEAUC

Table 3. Parameter ranges used in different combinations per task. The parameters were adapted in the process shown in figure 3.

Parameter	Range
Length of moving window for averaging	None—60 epochs on selected features
Estimators/trees	80–500
Min. sample leave	2–5
Split criterion	Gini or entropy
Probability threshold for majority class	0.65–0.9
Polynomial transformation	See table 4
Used features	See table 4

states were added to the classification task. A performance overview of the following κ values can be found in figure 8 and table 4.

Next, the CTW states are added to the classification task. Due to the low amount of wake but the similar physiological state (being awake, increased heart rate, increased number of noise), both states are merged. Adding CTW to be differentiated from AS and QS, the performance for both signal types, ECG and cECG, reduces. The classification using the ECG signal leads to a performance of κ 0.30 $\pm$ 0.12; 0.32. Using the cECG signal, the performance resulted in κ of 0.44 $\pm$ 0.21; 0.41. Trying to differentiate between AS, QS and IS based on the ECG signal, a performance of κ 0.23 $\pm$ 0.15; 0.29 was achieved. Using the cECG signal here, the performance reached κ 0.36 $\pm$ 0.12 0.34.

When differentiating between all states, using the ECG a performance of κ 0.30 $\pm$ 0.18; 0.23 was achieved. Using the cECG the performance reached κ 0.34 $\pm$ 0.17; 0.33. Each classification used different input parameter values found via grid search and manual extraction. As those parameters are specific for the here presented data,

Table 4. Mean kappa performance and used features per dataset and per class-set. Feature index used as in table 1. Features in bold were transformed.

Task	Signal	Selected features	Performance [κ]
AS|QS	ECG	0,..,5,7,11,..,16,18, 19,**21**,..,25,30,32,33	0.42 ± 0.26
	cECG	0,..,5,7,11,12,14,17, 19,20,22,24,..,29,32	0.56 ± 0.20
AS|QS|IS	ECG	0,..,**21**,22,**23**,**24**,..,33	0.23 ± 0.15
	cECG	**1**,2,3,6,7,10,..,19, 21,..,29,**32**,33	0.36 ± 0.12
AS|QS|CTW	ECG	**0**,..,10,12,**13**..,18,21, 23,..,26,30,..31,**32**,33	0.30 ± 0.12
	cECG	0,**1**,**2**,5,..,9,11..,18, 20,..,29,32,33	0.44 ± 0.21
AS|QS|IS|CTW	ECG	0,4,..,29,32	0.30 ± 0.18
	cECG	1,2,**4**,**5**,6,7,10,11,12,14,**15**,..,**18**,..,**27**,28,29,**32**	0.34 ± 0.17

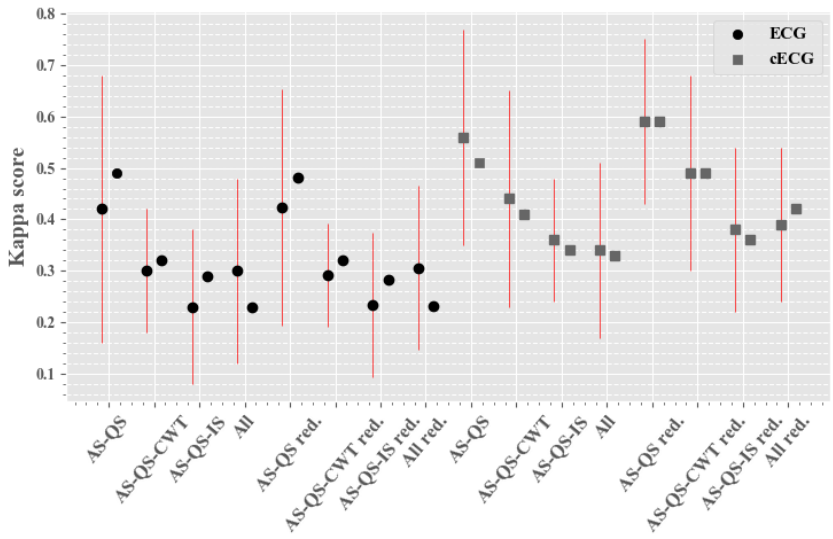

Figure 8. Classification performance for the ECG and the cECG for each class-set with mean $\kappa \pm$ std and cumulative pooled κ. Also displayed is the performance using the reduced set, were two outlying patients were excluded. The results from the reduced sets are labelled with the extension *red*.

it was chosen not to present the single parameter values. The general concept, use of the parameters, and selection method were described preceding.

To determine the maximum performance two main outliers were dropped. The two outliers were the patient's lowest in age and weight. Both patients were ventilated with continuous positive airway pressure (CPAP). For the ECG signal, dropping the outliers did not bring any improvements except that the standard derivation stabilized on a lower value for the kappa per patient. When dropping two outliers for the cECG, increased performance could be measured. The mean increase was κ of +0.04 for the mean kappa performance and κ of +0.07 for the cumulative performance measure. Reaching κ 0.59 ± 0.16; 0.59 for AS and QS classification, κ of 0.49 ± 0.19; 0.49 for AS, QS, and CTW classification, κ 0.38 ± 0.16; 0.36 for AS, QS, IS classification and κ of 0.39 ± 0.14; 0.42 for all state classification.

The feature sets used for the different class-set can be seen in table 4. Those feature sets were used for all patient data and reduced dataset with excluded outliers.

4. Discussion

4.1. Annotations

The ground truth annotations were also based on ECG signals but mainly video observations were used for the manual sleep state annotations. Thereby, features were included in the annotations which are not captured by the ECG and subsequent cannot be analyzed creating a general difference between the ground truth and the ECG/HRV approach. Unfortunately, the videos were not always of high quality regarding patient visibility. In few cases, the movements and breathing could only be seen passively through a moving blanket, and the annotators had to rely solely on the vital signs, knowledge of the sleep cycle and their experience with the patient. Therefore, some annotations might be slightly distorted compared to the actual state. This circumstance is only noted for completion as in our appraisal this slight distortion does not majorly affect the classification performance.

Additionally, in the annotations, sections were found with very unlikely sleep state sequences. This led us to believe that the ground truth is to some extent not representing the actual situation and possibly leading to a minor decreased performance of the classifier.

4.2. Features

Identifying the features that are most important for the classification, it can be seen that separating the classes using the ECG data the standard time domain and frequency domain features are prevailing. Low and high-frequency features are prominent in almost all state-group combinations. Low frequencies are associated with baroreflex activities and the high frequencies link to the parasympathetic system and respiratory activities. In contrast, the dominant cECG features represent more the general signal structure and account for noise and movement artifacts. We believe that this is due to the cECG being more sensitive to movements and therefore picking up the difference between small jitters and no movement expressive for AS and QS. CTW is mainly represented by movement artifacts and also external noise from nurse handling making this state more susceptive for noise/movement based features.

It is known that respiration is a distinct indicator for sleep states. The respiratory sinus arrhythmia, known in adults, is not pronounced in preterm infants. Therefore, breathing might only be picked up through breathing motion artifacts rather than modulated ECG signals as in adults (Fonseca *et al* 2015). High-frequency features, which are linked to breathing activities, are seen as essential features with the ECG while in the cECG, the motion features dominate as they possibly resemble breathing stronger than the HRV frequencies. That breathing is picked up by the cECG system is quite likely as it is highly sensitive to movement artifacts.

From the literature, we assumed that pDEC and SDDEC would have more impact on the classification, specifically for discriminating CWT (Kommers *et al* 2017). Nevertheless, as their impact increases mainly in stressful periods, they are possibly distorted by the noise which naturally appears during stressful periods (e.g. caretaking). Possibly, they will be more prominent in the ECG when more data is available, and distinct deceleration patterns can emerge.

4.3. Classifiers

Trying different classifiers, the RF came out on top each time. Classifier benchmarks appear to be not representative in this case due to the low numbers of analyzed patients. However, we will not to attempt to discuss the probable causes of RF being superior in this case. In our earlier publication, we used the RBF kernel SVM successfully. As SVM, in general, performs very well on binary class problems, it was here confronted with multiclass classification where RF generally is suited better. SVM is known to be particularly good dealing with outliers. As the data had two patients who constituted as outliers, SVM would be expected to perform better in this case. Nevertheless, RF can also handle outliers and in addition, tends to generalize better, creating a more stable model due to the randomization of data samples, especially on non-sparse data. Important is also that the RF needs much less parameter tuning than a kernel SVM. With the here chosen approach, many parameters fed into the system which increased the calculation time. Hence, a more exhaustive parameter search for the SVM was not feasible which possibly lead to suboptimal parameter settings and ergo performance. The same arguments hold for the RF versus the GB tree classifier, which needs well-tuned parameters to outperform a random forest approach. Choosing the RF decreases the chance of overfitting due to the parameter tuning to the dataset at hand. In contrast to an RF classifier, a GB tree learner aims towards reducing the bias introduced by the use of shallow trees. For the RF the bias is as high as the bias of the single sample trees and cannot be reduced. As our train and test sets are created based on patients, the bias is already significantly reduced limiting the impact of the GB approach. Another benefit of using RF rather than SVM and GB is the robustness of RF against overfitting (Breiman 2001) by the random selection process. Overfitting can further be reduced by limiting the tree depth using min sample leaf and min sample split which are adding more regularization. The ERF performs similarly to the RF classifier. The waiving of bagging increases the performance of ERF over RF for large datasets. Nevertheless, here the RF still outperformed the ERF slightly. For larger datasets we would recommend to look further into ERF also because of the reduced calculation time compared to RF.

4.4. Classifications

Coming back to the classification results per state and the underlying rationale the central sleep states AS and QS are mostly defined by the para/-sympathetic nervous system activities and its response to external triggers. When adding CTW or IS, the differentiation becomes more difficult, especially with such a small amount of training data. It can be seen in the confusion matrices that IS is falsely classified mostly as AS and some QS but least as CTW. Logically, those misclassifications rise from the IS definition as an in-between state. IS carries elements of all states and is not well defined regarding vital sign and observation boundaries. CWT is mainly misclassified as state QS or state IS. Those misclassifications seem to break ranks as it is expected that via the movement elements CWT would mainly be misclassified as AS. This is due to several reasons. First, CWT has the least training data

which results in decreased pattern recognition. Additionally, caretaking induces hugely varying patterns making it more difficult to train and re-identify on specific patterns. In the videos, it could be identified that between handling episodes during caretaking the patient would show no signs of movement possibly leading to a false classification as QS. Last, misclassification is linked to the way the predicted labels where joined under probability threshold rules. AS and QS state predictions are determined to focus on the minority class. Next, IS and CTW is joined under probability ruling which can lead to those misclassifications. The miss-prediction of states in both steps means that there might be further room for optimization by adjusting the probability thresholds even further while they are already tuned in a fine balance accepting this tradeoff.

As the dataset is limited in the amount of data available for training, we removed two outliers to determine the performance on a rather regular patient group, even though regular is a rather inept term regarding preterm infants. We noticed that removing those outliers actually only impacted with the cECG. While investigating into this matter, it was confirmed that the video quality was decent and therefore the video annotations could not pose the core problem. Most prominent is the fact, that the two preterm infants with a poor performance where both 27 wk GA in contrast to the others with 29 wk GA and older. One of the two patients also had a very low birth weight of 755 g leading probably to very unstable general conditions. Also, the preterm infants with poor performance which were partly excluded had CPAP devices in use. This could indicate that the use of CPAP obscured the SNS control of the breathing. As the effect through the included outliers could only be seen in the cECG and not the classic ECG, sleep state alteration through CPAP can most likely be suspended. It seems more likely that the CPAP masks the actual breathing pattern, substantiating, again, the importance of breathing pattern analysis for sleep staging.

With this small amount of data, the results cannot generally speak for preterm infant sleep classification but can draw a picture of the possibilities using the cECG and shows the impact of different factors on the classification such as ECG-features resembling breathing rhythm. We assume that using sole ECG signals will be on par with cECG when more data is available. Additional, data will produce a more stable model for classification. The use of a *min sample leave* value of below ten shows that this model is slightly tailored to this particular dataset, not leading to overfitting but having little training data, single epoch outlier have a more dominant impact on the classification performance.

Also, throughout this analysis, we included and excluded a variety of parameters impacting the classification performance. This manually backward forward approach is predestinated for missing the superior classification paths by the wide range of tuning factors. Therefore, this approach would be better suited for a deep learning strategy, were complex and nonlinear cross dependencies are all taken into account. Ansari *et al* (2018) showed the successful classification of QS in preterm infants using deep learning algorithms based on EEG signals. Unfortunately, for this preliminary study with limited data points, a deep learning strategy was outside the possible.

In summary, the bi-state classification is possible, especially for the majority classes AS and QS which are the most important classes for an automated neuronal maturation monitoring. All state classification is at this point not feasible. More data is needed for a stable all state classification. All state classification would be necessary for a holistic view on the patient's sleep and improved predictions of sleep cycles.

5. Conclusion

It was shown that cECG signals can be used for preterm infant sleep staging, separating AS and QS in an acceptable manner. However, analysis of all states including IS, caretaking, and wake becomes challenging. Movement alone is a strong indicator/separator for preterm infant sleep states. Incorporating features reflecting movement may help to detect sleep-associated respiratory activities, as there exists a strong connection between breathing pattern and preterm infant sleep states. As this study included only a small number of patients, further investigations for more generalized and improved ECG based sleep staging algorithms should be started.

Acknowledgments

The authors declare that there are no conflicts of interest. There is also no funding to declare.

ORCID iDs

Jan Werth https://orcid.org/0000-0002-9172-1385
Ronald M Aarts https://orcid.org/0000-0003-3194-0700
Xi Long https://orcid.org/0000-0001-9505-1270

References

Ansari A, De Wel O, Lavanga M, Caicedo A, Dereymaeker A, Jansen K, Vervisch J, De Vos M, Naulaers G and Van Huffel S 2018 Quiet sleep detection in preterm infants using deep convolutional neural networks *J. Neural Eng.* **15** 066006

Atallah L, Serteyn A, Meftah M, Schellekens M, Vullings R, Bergmans J W M, Osagiator A and Oetomo S B 2014 Unobtrusive ECG monitoring in the NICU using a capacitive sensing array *Physiol. Meas.* **35** 895–913

Becker D E 2006 Fundamentals of electrocardiography interpretation *Anesth. Prog.* **53** 53–64

Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32

Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 SMOTE: synthetic minority over-sampling technique *J. Artif. Intell. Res.* **16** 321–57

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

Duda R O, Hart P E and Stork D G 2000 *Pattern Classification* (NewYork: Wiley) ISBN:0471056693

Fonseca P, Aarts R M, Long X, Rolink J and Leonhardt S 2015 Estimating actigraphy from motion artifacts in ECG and respiratory effort signals *Physiol. Meas.* **37** 67–82

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

Gordon D H 1975 Triboelectric Interference in the ECG *IEEE Trans. Biomed. Eng.* **BME-22** 252–5

Gruetzmann A, Hansen S and Müller J 2007 Novel dry electrodes for ECG monitoring *Physiol. Meas.* **28** 1375–90

He H, Bai Y, Garcia E A and Li S 2008 ADASYN: adaptive synthetic sampling approach for imbalanced learning *2008 IEEE Int. Joint Conf. on Neural Networks (IEEE World Congress on Computational Intelligence)* pp 1322–8

Isler J R, Thai T, Myers M M and Fifer W P 2016 An automated method for coding sleep states in human infants based on respiratory rate variability *Dev. Psychobiol.* **58** 1108–15

Kommers D, Joshi R, van Pul C, Atallah L, Feijs L, Oei G, Bambang Oetomo S and Andriessen P 2017 Features of heart rate variability capture regulatory changes during kangaroo care in preterm infants *J. Pediatr.* **182** 92–98.e1

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

Lake D E 2011 Improved entropy rate estimation in physiological data *Proc. Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society EMBS* pp 1463–6

Lucchini M, Fifer W P, Sahni R and Signorini M G 2016 Novel heart rate parameters for the assessment of autonomic nervous system function in premature infants *Physiol. Meas.* **37** 1436–46

McHugh M L 2012 Interrater reliability: the kappa statistic *Biochem. Med.* **22** 276–82

Natekin A and Knoll A 2013 Gradient boosting machines, a tutorial *Front. Neurorobot.* **7** 21

Otte R and Long X 2019 A framework for observational coding of sleep in infants (in preparation)

Pedregosa F *et al* 2011 Scikit-learn: machine learning in Python *J. Mach. Learn. Res.* **12** 2825–30 (http://dl.acm.org/citation.cfm?id=1953048.2078195)

Prechtl H F R 1974 The behavioural states of the newborn infant (a review) *Brain Res.* **76** 185–212

Rooijakkers M J, Rabotti C, Oei S G and Mischi M 2012 Low-complexity R-peak detection for ambulatory fetal monitoring *Physiol. Meas.* **33** 1135–50

Ruf T 1999 The Lomb–Scargle periodogram in biological rhythm research: analysis of incomplete and unequally spaced time-series *Biol. Rhythm Res.* **30** 178–201

Scher M S, Turnbull J, Loparo K and Johnson M W 2005 Automated state analyses: proposed applications to neonatal neurointensive care *J. Clin. Neurophysiol.* **22** 256–70

Serteyn A, Vullings R, Meftah M and Bergmans J W M 2015 Motion artifacts in capacitive ECG measurements: reducing the combined effect of DC voltages and capacitance changes using an injection signal *IEEE Trans. Biomed. Eng.* **62** 264–73

Veen J, Meftah M, Lambert N, De B B M, Feddes B, Gourmelon L, Rietman R and Husen A 2011 Electro-physiological measurement with reduced motion artifacts *US Patent* US9603542 (https://patents.google.com/patent/US9603542)

Vullings R, Peters C H L, Hermans M J M, Wijn P F F, Oei S G and Bergmans J W M 2010 A robust physiology-based source separation method for QRS detection in low amplitude fetal ECG recordings *Physiol. Meas.* **31** 935–51

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

Werth J, Long X, Zwartkruis-Pelgrim E, Niemarkt H, Chen W, Aarts R M and Andriessen P 2017b Unobtrusive assessment of neonatal sleep state based on heart rate variability retrieved from electrocardiography used for regular patient monitoring *Early Hum. Dev.* **113** 104–13

Wijshoff R, Mischi M and Aarts R M 2017 Reduction of periodic motion artifacts in photoplethysmography *IEEE Trans. Biomed. Eng.* **64** 196–207

Yentes J M, Hunt N, Schmid K K, Kaipust J P, McGrath D and Stergiou N 2013 The appropriate use of approximate entropy and sample entropy with short data sets *Ann. Biomed. Eng.* **41** 349–65