Deep learning approach for ECG-based automatic sleep state classification in preterm infants

Citation for published version (APA):
Werth, J., Radha, M., Andriessen, P., Aarts, R., & Long, X. (2020). Deep learning approach for ECG-based automatic sleep state classification in preterm infants. *Biomedical Signal Processing and Control*, *56*, Article 101663. https://doi.org/10.1016/j.bspc.2019.101663

Document license:
TAVERNE

DOI:
10.1016/j.bspc.2019.101663

Document status and date:
Published: 01/02/2020

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Biomedical Signal Processing and Control 56 (2020) 101663

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Deep learning approach for ECG-based automatic sleep state classification in preterm infants

Jan Werth[a,*], Mustafa Radha[b], Peter Andriessen[c], Ronald M. Aarts[a,b], Xi Long[a,b,**]

[a] Department of Electrical Engineering, University of Technology Eindhoven, Eindhoven, the Netherlands
[b] Philips Research, Eindhoven, the Netherlands
[c] Paediatric Department, Máxima Medical Center, Veldhoven, the Netherlands

ARTICLE INFO

Article history:
Received 18 December 2018
Received in revised form 7 April 2019
Accepted 17 August 2019

Keywords:
Preterm infants
Sleep
Automated sleep state classification
Electrocardiography
Deep learning
Recurrent neural networks
Long short-term memory
Gated recurrent unit

ABSTRACT

Preterm infant neuronal development is related to the distribution of their sleep states. The distribution changes throughout development. Automated sleep state monitoring can become a powerful aid for development monitoring in preterm infants. Three datasets including 34 preterm infants and a total of 18,018 30 s manually annotated sleep intervals (sleep-epochs) were analyzed in this study. The annotation of sleep states includes active sleep, quiet sleep, intermediate sleep, wake, and caretaking. Four different recurrent neuronal network architectures were compared for two-state, three-state, and all-state analysis. A sequential network was used to compare long- and short-term memory and gated recurrent unit models. The other network architectures were based on the popular ResNet and ResNext architectures utilizing residual connection for more depth. The most essential sleep states, active and quiet sleep, could be separated with a kappa of 0.43 ± 0.08. Quiet versus caretaking and wake showed a kappa of 0.44 ± 0.01. The three state classifications of active versus quiet versus intermediate sleep resulted in a kappa of 0.35 ± 0.07 and active versus quiet versus wake and caretaking resulted in a kappa of 0.33 ± 0.04. The all-state classification was underperforming with probably due to difficulty in separating subtle differences between all states and a lack of sufficient training data for the minority classes.

1. Introduction

Preterm infant sleep shows several distinct sleep states. They are defined mainly as active sleep (AS), quiet sleep (QS), and wake. In very preterm infants sleep states may be rudimentary and show transitionary shifts from one to the other, often with patterns of both, AS and QS, intermediate or undetermined (IS) states [1]. AS is often compared to the adult rapid eye movement (REM) sleep states because it shows similar increased neural activity [2–4], nevertheless, the role of preterm infant sleep states seems to be different. It is assumed that the sleep, the sleep states and the sleep cycles of the fetus, preterm and term infants all play an essential role in the sensory and cortical development [5–7]. Initially, AS is providing stimulation to the newborn brain in a sensory-reduced environment triggering the development of brain regions with reduced sensory input [2]. Furthermore, during AS the development, integration, and alignment of specific neural tasks/regions into the cortex structure is taking place. During QS, it is reasonably assumed that developmental errors are corrected, and reorganizations are conducted with the use of increased brain plasticity. During QS the parasympathetic nerve activity is dominant, blood pressure and heart rate are lowered and therefore often seen as the resting and re-energizing state [5,8]. AS sleep is dominating the sleep cycle of preterm infants with about 80% of the total sleep time at early gestational birth. QS is seen as the minority state with about 18% of the total sleep time. The distribution changes in the course of development with decreased AS and increased QS [9–11]. The states can be observed by differences in many electrophysiological signals, such as vital signs (e.g., heart rate and respiratory rate), movement, and electroencephalographic (EEG) activity. During active sleep, increased cardio-respiratory activity and increased motor activity with sporadic eye movements can be observed. During QS all cardio-respiratory activities are lowered in amplitude and dynamic range [1,11,12].

From our experience, at present, manual, sporadic sleep annotation is still the clinical standard for sleep classification and analysis. Continuous and automated monitoring of sleep in newborns may

* Corresponding author at: Carl Orff Strasse 37, 55127 Mainz, Germany.
** Corresponding author at: Philips Research, HTC 34, 5656 AE Eindhoven, the Netherlands.
E-mail addresses: werth.jan@gmail.com (J. Werth), xi.long@philips.com (X. Long).

https://doi.org/10.1016/j.bspc.2019.101663
1746-8094/© 2019 Published by Elsevier Ltd.

provide clinical decision support by optimizing the workflow of the caretakers, avoiding disruption of AS/QS and, most importantly, to better safeguard the preterm infant's developmental process. To provide such a monitoring system in a neonatal intensive care unit (NICU) setting, no additional sensor(s) should be introduced. Therefore, the system should concentrate on already existing, continuously monitored parameters such as electrocardiography (ECG). To date, the most successful approaches regarding preterm infant sleep monitoring are using EEG signal analysis [13,14]. Unfortunately, EEG and even the reduced aEEG have to introduce additional sensors to the routinely used monitoring solutions, such as ECG, to enable sleep monitoring. As the preterm infant skin is highly sensitive with only three layers thick epidermis and almost no outer protective skin layer [1,15,16], additional electrodes should be avoided and EEG/aEEG is not used for a regular and continuous solution at this point. To overcome these limitations motivated to investigate ECG regarding preterm infant sleep monitoring. Further motivation for the use of ECG as signal modality is the fact that ECG can also be obtained unobtrusively, with, e.g., capacitive ECG [17], and could thereby be utilized for a future non-contact sleep monitoring solution.

1.1. Related work

Machine learning opens up the possibility to create an automated algorithm based only on ECG. In the adult sleep research, machine learning algorithms have already been successfully used [18,19]. Radha et al. [20] compared several machine learning methods such as random forest and ensemble support vector machine (SVM) to show promising real-time EEG sleep analysis. Further, full polysomnographic analysis (PSG) [21,22] and unobtrusive actigraphy methods [23] were investigated for adult sleep state separation.

Sleep analysis for preterm infants is more difficult as the states are less distinguishable. Nevertheless, machine learning for automated EEG [14] analysis, ECG for sleep vs. wake [24] and heart rate variability (HRV) for sleep states analysis [25] have been investigated, demonstrating the potential of machine learning for preterm infant sleep state classification.

A step further in the automated analysis is the use of artificial neural networks (ANN), and respectively deep learning comes naturally [26]. The advantage of ANNs over more traditional machine learning is that ANNs can learn richer representations and model complex non-linear relationships. This is of specific importance when dealing with human data as there appear many nonlinear, and complex relationships between the in- and output. In addition, ANNs are easy to change in their architecture to adapt to the complexity level of the problem at hand. Also, ANNs are non-parametric models that have the advantage over parametric models that any given input distribution can be learned without prior knowledge. This comes especially in handy when multi-source input features are used with unknown or different distributions. Furthermore, convolution derived features could, in theory, be superior in separating sleep stages than handcrafted features based on ad-hoc decisions such as thresholds, filter cutoffs or similar approaches. The latest development in ANN originated specific units for time series analysis in recurrent neural networks (RNN) [26], the long short-term memory (LSTM) [27] and gated recurrent unit (GRU) [28]. Those are specifically designed to find patterns in time, such as sleep architecture, to generate improved performance. As the development focuses currently on ANNs, more groundbreaking renovations, also benefitting sleep analysis, might be expected in the future. The application of ANNs to the topic of sleep classification is on the rise. In 2017, Bishwal et al. [29] presented the annotation tool SLEEPNET using a large dataset to train a deep recurrent neural network (RNN) reaching human level annotation performance. During the same period, Chambon et al. [30] published the implementation of an algorithm that is independent of crafted features using convolution in combination with spatial filtering for classification. A similar approach was chosen by Supratak et al. [31] using an ensemble of convolutional neural networks (CNN) and RNN networks to be able to classify sleep from raw EEG data. At the beginning of 2018 Olesen et al. [32] presented an approach with an adapted model using transfer learning from the ResNet50 architecture. In the following, Sano et al. [33] used long- and short-term memory (LSTM) classifier to identify wake vs. sleep from multimodal data. One of the most recent publications on the topic from Radha et al. [34] used LSTM classifier to classify sleep from HRV features overcoming the temporal limits of non-temporal models. Also, they used transfer-learning to enable the utilization of other signals (here photoplethysmography) with the same trained model enabling different application areas.

Sleep classification in adults is well handled with deep learning, less studied in preterm infants. So far, Ansari et al. [13] are the first to implement a CNN network for preterm infant EEG signals successfully. This publication tries to investigate the possibility of using an RNN approach for the more difficult preterm infant sleep classification. We hypotheses that with deep learning algorithms which incorporate time domain analysis, such as with RNN architectures, preterm infant sleep states can be classified in an acceptable accuracy only based on ECG features. We predicate this hypothesis on the base that ANN networks are distinguished on analyzing highly complex systems which are influenced and dependent on multi-system factor interrelations. And preterm infant sleep is such a multi-factor influenced system. Generally, ANNs should outperform classic machine learning methods; however, based on limited data in this study we believe that our ANN approach will perform equally to previous machine learning methods but having overall greater potential.

2. Methods

2.1. Population

Datasets from three retrospective studies were combined. The dataset recordings have a timespan of several years in-between them. The infants were admitted to the 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, and written consent was given by the patient's parents. In those three retrospective studies, 34 (8, 9, 17) stable preterm infants were analyzed during 39 sessions. The preterm infants were born with a mean gestational age (GA) of 29 ± 2.1 weeks. They were studied at a mean postmenstrual age (PMA) of 33 ± 2.0 weeks. The patients had a mean birth weight of 1338 ± 473 g.

2.2. Data recordings

Vital signs recordings for all studies were performed with a Philips patient monitor (Intellivue MX 800, Germany) at a sampling frequency of 500 Hz (n = 32) or 250 Hz (n = 2). The 250-Hz data were interpolated to meet the 500 Hz.

Each preterm infant was also video-recorded. Videos were either recorded of the face or the total body view. The used cameras were standard, medium resolution, greyscale devices.

2.3. Annotations

Per dataset, two trained observers annotated the data based on 30 s intervals (sleep-epochs) adhering the Prechtl system [35]. The observers used a reference ECG and respiration time series and video information for annotation. Vocalization, which is part

Table 1
State distribution per dataset in percent.

States	Dataset 1 [%]	Dataset 2 [%]	Dataset 3 [%]
Unknown	1,6	37,9	0,0
AS	65,4	47,7	46,0
QS	7,5	6,7	18,9
Wake	2,0	1,6	11,9
CT	8,4	0,0	0,0
IS	15,1	6,0	23,3

of the Prechtl system, could not be used due to the lack of audio recordings. They annotated the following states: AS, QS, IS, wake, caretaking, and unknown (unable to annotate). The more specific states from Prechtl active wake and quiet wake were merged into wake due to lack of data. The total duration of annotated data was 167 h (20,021 30 s intervals) with a mean duration per patient of 4.28 ± 1.5 h (513 ± 179 30-s intervals). The overall distribution of state was: AS: 51.45%, QS: 12.7%, IS: 16.5%, wake: 6.6%, caretaking: 2.2% and unknown: 10.5%. Subtracting the unknown epochs, a total amount of around 18,018 30 s intervals were left for analysis. The detailed distribution of all trials can be found in Table 1. The median inter-rater variability lied at 0.8 ± 0.1 for AS and 0.6 ± 0.1 for QS. The overall median inter-rater variability was 0.7 ± 0.1.

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) to equalize for the low amount of data from each state.

2.4. ECG R-peak detection

The R-peak detection algorithm of Wijshoff et al. [36] was used to determine the NN intervals and the resulting HRV signal. To determine the steepest ascent and descent of the QR and RS slopes they calculated the first derivative of the ECG signal. Then the peaks in the QRS complex were detected with a variable threshold. By interpolation around the detected peaks, they verified that the position of the peak is at the real max. This sub-peak detection assured that there is no shift from the real peak due to off sampling.

2.5. Features

For each dataset 47 features from HRV, ECG, and patient information were created. The features were calculated based on 30 s intervals. The HRV features include the time, frequency and nonlinear domain. The ECG features were calculated in the time and nonlinear domain, while the ECG derived respiration (EDR) features were calculated in the frequency and nonlinear domain. For HRV and EDR the signals are fundamentally non-equidistant in time. The Lomb-Scargle algorithm [37] was used to generate the frequency spectrum as resampling for classic Fourier transformation would have introduced extra parameters.

As the respiratory sinus arrhythmia (RSA) and cardiorespiratory coupling is not very pronounced in preterm infant and can only be seen in more mature infants [38,39], Joshi et al. [40] confirmed very recently that coupling between heart rate decelerations and accelerations exist in preterm infants but not vice versa. They assume that RSA in preterm infants is not present mainly due to insufficient breathing depth. The respiration can consequently not be determined from the RSA but rather via superimposed chest movement on the ECG signal. Therefore, the EDR signal was calculated using the ECG envelope. In the frequency domain, the frequency band was limited to max 1.1 Hz (66 breaths per minute) and min 0.3 Hz (18 breaths per minute), which is described in the literature as the min and max respiration rates of preterm infants [41].

The frequency bands were then separated into high (1.1–0.84 Hz), medium (0.84–0.56 Hz), and low (0.56–0.3 Hz) bands.

From the patient information data, gestational age (GA), age at measurement (CA), and birth weight (BW) were taken. To gain the timespan between birth and data recording, CA and GA were subtracted from each other. All this information was combined into a stability score. This score would indicate either an unstable, medium or stable patient condition.

All features are listed in Table 2. The normalized features with mean zero and standard derivation of one were combined into 3D tensors which were fed as input into the deep learning models.

2.6. Preprocessing for deep learning

For the classification of the preterm infant sleep states, the neural network API Keras [42] was used with TensorFlow [43] backend. For sleep state classification, time series analysis was used. Therefore the input was cast in the form of a 3D time series tensor [samples, time step, features]. After testing, the time step was chosen as the total length of one recording session with the batch size set to 1. Thereby, long and short-term patterns can be recognized. To achieve uniform length, the tensors were padded to the length of the most extended session. Later, a masking layer and sample weight distribution of zero for the padded values was used to prohibit the padded values to influence the learning process. The data was separated into train and validations sets to exclude significant bias. The data was split per patient to reduce the bias for the train and validation process further. The split per patients was set to 70% training data and 30% validation data (ceiled). A 3-fold cross-validation process was used to ensure the proper generalization of the model.

As we have a majority (AS) and minority classes (QS, IS, CW) a class weight has to be calculated to balance this unequal class distribution. The sample weight was calculated depending on the sleep state and normalized to the majority class. Sample weight was used instead of class weight as class weight is converted to sample weights on the Keras backend. Using sample_weight_mode temporal, sample weight fulfills the class weight task and can as well be used for masking-padded-values with a sample weight of 0. So far, this can only be used for smaller datasets as sample_weight_mode does not work currently for the function fit-generator.

2.7. Classification models

Four different model types were compared: a deep residual model, a wide residual model, a wide residual model using transfer learning from sequential models, and the sequential models as standalone architectures.

The base residual architecture itself was adapted from the residual architectures ResNet [44] and ResNext [45]. Both approaches tackle the problem that an increase of model depth creates a sudden and rapid decrease of accuracy, which is not caused by overfitting but rather shattered gradients [46]. The shattered gradient appears in none residual networks appear, with large depth, like white noise. Both networks combine multiple sequential models to one larger model. The sequential models thereby fit a residual map which is easier to optimize than a larger model. The connections between the sequential models are ensured with skip layers performing identity mapping and feeding the output of a sequential model block into the next block (see Figs. 1 and 2). This process enables large networks with rather low complexity. As preterm infant sleep is a highly complex problem due to the many in- and exogenous influences on the sleep and ANS, residual architecture approaches might be able to constitute and handle this complexity. The ResNet and ResNext architectures were both developed for

Table 2
Overview of the used ECG and HRV features for classification.

NR	Feature [unit]	Description
0	BpE	Beats per Epoch/mean 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 the 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 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 defined window size.
19	LFnorm [%]	LF power in normalized units LF/(Total Power-VLF) × 100
20	LF/HF [n.u.]	Ratio LF/HF
21	pHF1 [ms^2]	The power of the high-frequency band between 0.4–0.7 Hz
22	pHF1norm [%]	pHF1 power in normalized units pHF1/(Total Power-VLF) × 100
23	TotPow [ms^2]	Total power or variance of NN intervals of defined window size.
24	pHF2 [ms^2]	The power of the high-frequency band between 0.7–1.5 Hz
25	pHF2norm [%]	pHF2 power in normalized units pHF2/(Total Power-VLF) × 100
26	VLF [ms^2]	The power of the very low-frequency band between 0.003–0.04 Hz of defined window size.
27,28	SE, 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
34	HF_R	The power of the high-frequency band of the respiration signal between 0.48–1.1 Hz of defined window size.
35	HFnorm_R	HF respiration power in normalized units. HF/(TotPow_R-LF_R) × 100 of the respiration.
36	MF_R	The power of the medium frequency band of the respiration signal between 0.56–0.84 Hz of defined window size.
37	MFnorm_R	MF power in normalized units of the respiration. MLF_R/(TotPow_R-LF_R) × 100
38	LF_R	The power of the low-frequency band of the respiration signal between 0.56–0.3 Hz of defined window size.
39	LFnorm_R	LF power in normalized units of the respiration. LF_R/(TotPow_R) × 100.
40	LF_R/HF_R	The ratio between the low and high respiration spectrum. LF_R/HF_R
41	MF_R/HF_R	The ratio between medium and high respiration spectrum. MF_R/HF_R
42	TotPow_R	The total power of the respiration frequency spectrum.
43	Age difference	Difference between age at birth and age at measurement.
44	Birthweight	Weight at time of birth
45	GA	Gestational age. Age at birth calculated from the last gestation.
46	CA	Conceptional age. Age at time of measurement

image classification and object detection. Therefore, they are using mainly Relu activated CNN layers coupled with dropout and pooling layers. In this manuscript, the CNN and pooling layers were replaced with recurrent layers to capture patterns and connections in time series data rather than in image data.

As both approaches are similar in the core residual idea and compare comparable to each other in object classification tasks with a slight advantage in floating point operations per second (FLOPS), speed, and error rate of the ResNext architecture. As it cannot be directly deducted how they would compare in a time series analysis task and in this specific setting, a comparison is appropriate.

Deep residual model: the deep residual network is made of an initiation block followed by five residual blocks of five connected GRU layers (Fig. 2). Each connected block ends with a dropout [47] and dense layer. The architecture is finished with a softmax activated dense layer. The initiation block consists of first a masking layer, which is needed as the data is padded. The masking layer is followed by a 1/2 dropout layer connected to a dense layer which, both combined, function as a feature selection phase. This combination first randomly reduces the input nodes trailed by reducing the dimensional space that forces the focus on the most distinguishing input information. This also can be seen as the replacement for the pooling layer used in the classic ResNet and ResNext architectures for reduction. The last layer of the initiation is a batch normalization to avoid vanishing/exploding gradients by the scale of backpropagated weights.

Wide residual model: the wide residual model (Fig. 1) uses three parallel paths preluded by an initiation block. The initiation block is of the same structure as in the deep residual model. The parallel paths are each made of blocks from bi-directional GRU layers. Each GRU layer uses dropout and recurrent dropout to minimize overfitting. As direct regularization of the L2-norm a kernel constrain is used over all axes. Direct kernel constraints work well in combination with dropout [47]. Each path has two of the described blocks which are connected via skip layers in the same way as with the deep model. All parallel paths are concatenated at the end leading to a dense layer with softmax activation to achieve final state predictions.

Wide residual model with transfer learning: transfer learning uses pre-learned information, resembled in the weights which are fixed in a model architecture. The fixed weights adds to the training process without being changed during the training. Thereby, pre-learned information can be used to reduce the overall learning computation effort or improve the learning process by adding specific information. In image classification or object recognition, transfer learning is used to fix earlier learned universal information of shapes in images, e.g., general shapes in a face such as lines and edges. Later layers learn more complex compositions of such general shapes for a specific set of images, e.g., the composition of chimpanzee faces.

Here the same wide residual model architecture was used, but additional paths were added. The loaded weights of the pre-trained models where fixed into those additional paths Fig. 3. The pre-trained models were trained on bi-class problems, always training two classes versus each other, learning the specific differences between only those classes. In the concatenation step, they are then used for decision making. To avoid any bias, the bi-models

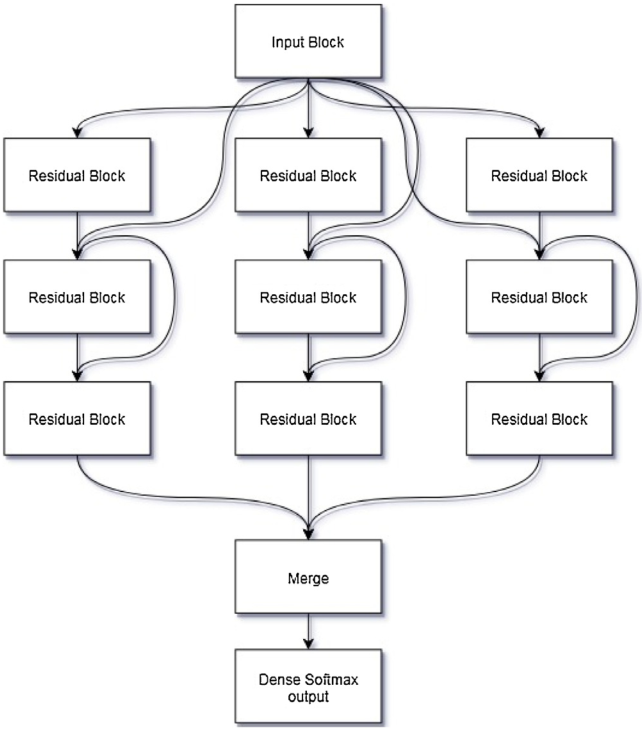

Fig. 1. Exemplary Wide residual model structure.
Wide residual model with Initiation block of masking layer, dropout layer, and following dense layer. Afterward, the architecture is split into three paths, where each path consists of connected bi-directional gated recurrent unit layers which are later concatenated again. The layers are connected with skip connections to help simply fining the network optimization. Each path uses different hidden units to incorporate more and less complex relations.

are trained separately only on a fragment of the data, which is later not used for further training.

Sequential models: the sequential models, which are also used for pre-training, have an initial masking, dropout and dense layer which is followed by four bidirectional GRU layers. The models are closed with another dropout layer and a dense, softmax activated layer. The hidden Units for the dense and GRU layers where set to 32. All other parameters followed the main transfer learning model. The model architecture was compared in performance and speed to the same architecture using bidirectional LSTM layers.

2.8. *Model parameters*

The wide and deep residual model, use a range of hidden units for the GRU layers to learn and model a wider range of complex non-linear relationships in the input data stream. The wide network uses a different hidden unit for each path ranging from 4 to 128 hidden units. The deep architecture increases the hidden units with each block from 32 to 256. The hidden units of the Dense Layers were set differently to accommodate the previously mentioned feature selection. The values ranged from 16 as roughly 1/3 of the input feature dimension and a power of 2, to a max of 47 representing the full input feature dimension.

To further handle the data imbalance, the earlier mentioned class weights were used in a weighted categorical cross entropy loss function to increase the misclassification gravity for minority misclassification. Therefore, the normalized weights multiplied with the loss function

$$L(s)_w = L(s) \cdot (1 - ||(s)) \quad (1)$$

where ll(s) is the apriori likelihood of s in the training data.

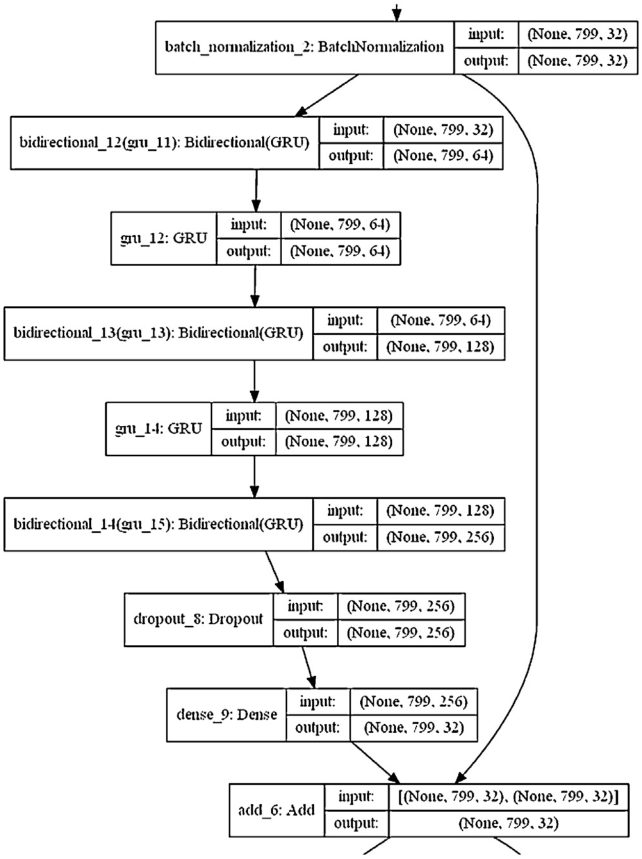

Fig. 2. Residual block of deep model.
An exemplary block from the deep residual model. Here gated recurrent units and bidirectional gated recurrent unit layers alternate each other with increasing hidden units (neurons). Thereby, the hidden units increase from 32 to 256 covering simpler to more complex feature-state connections.

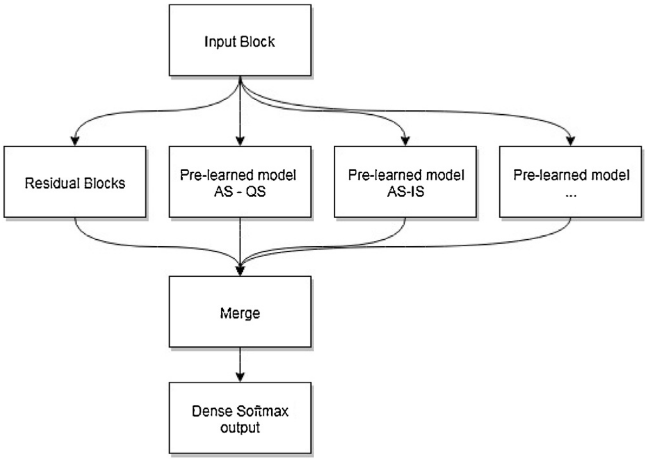

Fig. 3. Architecture of the residual model utilizing transfer earning.
The input block and the Residual block are the same as in Fig. 1. Also, pre-learned blocks of sequential architectures are added parallel and fixed until a various point. All weights are concatenated and added into a dense layer with softmax activation.

The *Sigmoid* function was selected as the activation function for each residual GRU/LSTM block. In the ResNet, and following residual architectures, the rectified linear unit (*ReLu*) function is used as the activation function. Using the ReLu activation function is very difficult and not advised with GRU/LSTM as it diverges, but mostly not necessary as the gating scheme of the GRU/LSTM itself deals with the vanishing gradients. Therefore, the *Sigmoid* activa-

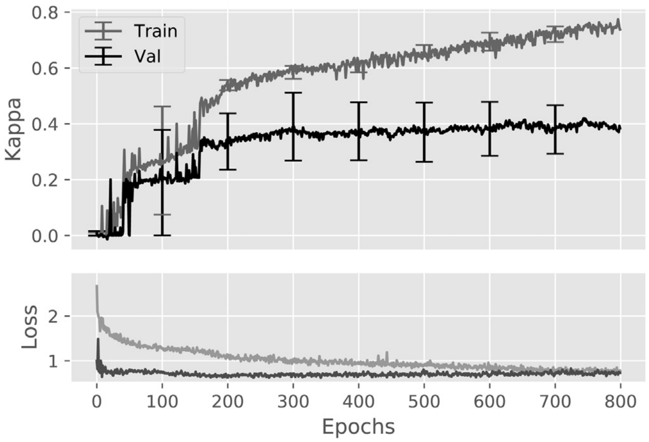

Fig. 4. Mean kappa and loss of active versus quiet sleep classification over epochs. Kappa and weighted categorical cross entropy loss over epochs using a sequential architecture. Initial dropout is 0.2, and dropout/recurrent dropout per bidirectional gated recurrent unit layer is 0.5. Kernel constraint per layer is set with max norm 0.3. Hidden units of 32 is used per layer.

tion, which is optimally designed for the GRU/LSTM structure, can be used.

For the optimization algorithm, the Adaptive Moment Estimation (Adam) optimizer [48] was chosen as Adam shows to be generally very effective while also removing the manual setting of the learning rate and learning rate decay. The Adam also was tested against other optimization algorithms where it stood out superior.

As already mentioned, the timestep (or lookback) was set to the longest recording session and the batch size to 1.

To avoid overfitting, dropout, L1 and, L2 regularizers were applied. The maximum dropout value was 0.6 before the results dropped off out of proportion. Combinations of L1 and L2 regularizations were implemented as kernel and activity regularizers in different places. L1 was implemented mainly as an additional feature selection in the first dense layer with a value between 0.0001 and 0.001. The L2 norm was used mainly on each LSTM/GRU layer. The L2 norm for direct kernel constraint was set to 0.3 on each layer.

3. Results

Here, three different types of classifications are presented. The two-state classification, which is a classification between two states. The three-state classification, which tries to separate three states against each other such as AS, QS, and IS. And all-state classification, which tries to separate all states from each other.

The two-state classification with a sequential architecture shows promising results for using GRU and LSTM layers. Both show similar mean results with a difference in performance of 0.01 ± 0.02. Due to slight faster modeling with the use of GRU layers, all results are presented using GRU layers.

The most robust performance is reached with the majority states AS and QS a mean kappa over the folds ranging from 0.43 ± 0.07 to 0.40 ± 0.06 (Fig. 4) and between QS and CTW with a mean kappa of 0.44 ± 0.01.

Then the combinations AS-IS and IS-CTW show similar results with 0.33 ± 0.03 and 0.32 ± 0.03. AS-CTW and QS-IS classification have the lowest performance of 0.28 ± 0.005 and 0.25 ± 0.03 (Table 3). Where Kappa score is defined as slight within 0–0.20, fair between 0.21–0.40, moderate with 0.41–0.60, substantial between 0.61–0.80, and as perfect within 0.81–1 [49].

The classification of three states shows results (Table 4) between two- and all state classifications (compare Table 5). The majority classes AS and QS were compared with the minority classes IS and CTW. The mean performances in Table 4 indicate that the Majority classes are better differentiable together with IS resulting in a mean kappa of 0.35 ± 0.07 rather than with CTW with a mean kappa of 0.33 ± 0.04.

Both deep and wide residual models show similar results for all state classification with a mean kappa of 0.30 ± 0.06 and 0.25 ± 0.02 (Table 5). Same as with the sequential model, the majority states AS and QS are separated best, followed by QS and CTW using wide and deep residual models. The overall performance is lower than for bi-class classification.

The use of transfer learning did not improve the performance even though the pertained models showed decent results. In contrary, it showed the lowest overall performance with a mean kappa of 0.13 ± 0.02 (Table 5) even though the performance of the prelearned models where acceptable (see Table 6).

Table 3
Mean performance for bi state classification using different model architectures. The Kappa value of two-state classifications of different model types. The residual models used a kernel L2 regularization of 0.01 and an activity L2 regularization of 0.001. Initial dropout is 0.2, and dropout/recurrent dropout per bidirectional gated recurrent units layer is 0.5. Kernel constrain per layer is set with max norm 0.3. Hidden units of 32 is used per layer. The sequential model used four long short-term memory or gated recurrent units bi-directional layers with sigmoid activation, Adam optimizer, and weighted categorical cross entropy loss function. The Kernel constrain was set to 3. 50% Initial and Recurrent dropout was used with both 0.5. Initial and Recurrent dropout was used. L2 activity and kernel regularization was set to 0.01 and 0.001 per layer.

State pairs	Residual Wide (κ ± std)	Residual deep (κ ± std)	Sequential (κ ± std)
AS-QS	0.37 ± 0.07	0.38 ± 0.04	0.43 ± 0.08
AS-IS	0.31 ± 0.03	0.30 ± 0.03	0.33 ± 0.03
AS-CTW	0.26 ± 0.005	0.25 ± 0.01	0.25 ± 0.03
QS-IS	0.28 ± 0.03	0.27 ± 0.007	0.28 ± 0.005
QS-CTW	0.39 ± 0.005	0.40 ± 0.001	0.44 ± 0.01
IS-CTW	0.30 ± 0.04	0.29 ± 0.03	0.32 ± 0.03

Table 4
Mean kappa performance of sequential models for three state analysis. Mean performance of three-state classification using a sigmoid activated model with four bi-directional gated recurrent units layers, Adam optimizer, weighted categorical cross entropy loss function. The Kernel constrain was set to 3. 50%. Initial and Recurrent dropout was used with 0.6 and 0.5. L2 kernel regularization per layer was set to 0.01.

State pairs	Kappa
AS-QS-IS	0.35 ± 0.07
AS-QS-CTW	0.33 ± 0.04

Table 5
Kappa performance on all state classification for different models. Mean kappa performance over different models using the same parameters as described in the other tables.

Model	Kappa
Deep Residual	0.30 ± 0.06
Wide Residual	0.25 ± 0.02
Sequential	0.25 ± 0.05
Wide Residual using Transfer learning	0.13 ± 0.02

Table 6
Kappa performance of sequential model architectures used for transfer learning. Kappa results without cross-validation using the same parameters as the sequential model in Table 3 utilizing bidirectional gated recurrent unit layers.

State pairs	Kappa
AS-QS	0.51
AS-IS	0.36
AS-CTW	0.27
QS-IS	0.29
QS-CTW	0.55
IS-CTW	0.38

4. Discussion

4.1. Features and patient demographics

The feature set was partly used in an earlier publication [17,25]. It was reused and adapted as it showed a good representation of the underlying processes of preterm infant sleep. Movement features were included as movement is generally a strong differentiator between sleep, wake, and caretaking. For future research improved measurement techniques for movement detection, as presented by Joshi et al. [50], or enhanced extraction methods for ECG would probably improve classification further. Additionally, patient information was added to the set as it was noticed in previous tests that outlying values of patients can seriously influence the performance of classification. Such values are in general not random but often occur at very young, immature preterm infants and/or with low birth weight. Also, it makes a huge difference in the development at which age a preterm infant was born and at which timespan after birth the data were recorded. If the measurement takes place at the same time with different GA at birth (or vice versa), the development state and consequently the feature appearance can look sufficiently different to influence the learning. The same hold for the birthweight. A heavier baby tends to be more stable. As the values of age, age difference, and weight are almost continuous data, they were categorized into a stability score between 1 and 3. With a significantly larger dataset either the values can be used directly, or a finer grid can be used to categorize the preterm infants. In previous tests, it was noticed that the use of respiration devices influence the classification performance. Unfortunately, it was not possible to gather this information for all patients.

4.2. Parameter choices

To avoid overfitting, various settings were applied with dropout, recurrent dropout, L1 and, L2 norm for activity (Ar) and kernel (Kr) regulation. Dropout and activity regularization had the most effect on overfitting. Different combinations of these regularizations were useful in reducing overfitting. Here only two are mentioned as examples. Either using overall lower dropout (e.g., 0.3) in combination with a stiffer Kr and Ar L2 norm (e.g., 0.01) without any other regularization in the initiation block helped fighting overfitting. Another variant is to use an L1 norm as kernel regularization in the first dense Layer to help with feature selection in the initiation block in combination with an overall dropout/recurrent dropout of 0.5–0.6 but no other Kr or Ar in the following layers. Choosing too high values for the dropout and/or regularization would lead to a drastic reduction in overall performance on the validation data. There were plenty of combinations that all resulted in reducing overfitting. Nevertheless, further investigation and proper comparison will go beyond the scope of this publication as overfitting was not the primary problem in this analysis.

The lookback was chosen as the total duration of one session for the LSTM/GRU as long-term sleep cycles can influence the overall learning process. As LSTMs/ GRUs can only learn the variations in time on the information of one batch, long-term patterns such as total sleep cycles or specific sleeping patterns need at least 30 min of data up to 70 min [51]. In regular cases, sleep states changes follow the pattern wake-AS-QS-AS-wake with IS patterns in between. Irregular patterns are for example wake-QS-(AS)-wake. This pattern is called a stress sleep pattern showing signs of the preterm infant's immediate need for rest. With more of such recorded patterns outside the norm, future research could try to detect anomalies in sleep cycle patterns to inform the responsible caretakers. Either regular or irregular cycle patterns cannot be learned with batches of insufficient length below at least one sleep cycle.

4.3. Classification performance

The classification between the majority classes AS and QS shows moderate performance with kappa 0.43 ± 0.07 and generally promising results. For a clinical monitoring device, this would not be an acceptable performance, but these results should be seen as proof of concept for an unobtrusive automated monitoring system. Especially with the earlier described benefits of ANNs, these results are a foundation that has a high chance to increase in performance, stability, and generalization with an increasing amount of accessible data. To put the result in context, it has to be beard in mind that the general interrater variability is relatively low in that specific population. In early studies about the reliability of polysomnography in term infants, the kappa score reached 0.68 [52]. For adult, the mean kappa score after the Rechtschaffen & Kales standard also reaches 0.68 and following the AASM standard 0.76 [53]. There is no overview accessible for preterm infants, but it is very likely to be lower than for term infants.

It is difficult to directly compare the here presented result to recent results from other groups as they are mostly based on EEG analysis, which is optimal for sleep state analysis as they directly represent the state of the autonomous nervous system. Koolen et al. [14] presented good results in 2017 using a support vector machine on EEG signals with 85% accuracy for AS and QS separation. Latest results using EEG signals were presented by Ansari et al. [13] using a CNN to classify QS and NonQS with a ROC-AUC of 0.92. Also, Dereymaeker et al. [54] were able to detect QS with an AUC of 0.97 based on EEG. The most similar approach from Isler et al. [55] used only respiration signals for successful classification of AS and QS with an agreement of 78% to 90% for AS and QS separation. In a previous publication of our group [25] AS and QS were separated with a ROC of 0.87 based only on ECG features. As more states were tried to separate, performance decreased. Fraiwan et al. [56] tried to separate AS, QS, and wake with a performance of 63% to 75% using also EEG analysis. Also, it has to be kept in mind that the here used kappa performance takes into account the unequal distributed states with increased expected accuracy.

Summarizing the results and methods used for sleep classification, Ansari et al. [13] can be considered the current state of the art for preterm infant sleep state analysis as they are using the latest analysis methods on EEG signals, which are the standard signals for sleep analysis, and achieved very high classification results for QS and nonQS states. Both of which are important for neural development monitoring.

Our moderate performance on AS and QS classification is also promising as the state distribution of AS and QS is one of the primary indicators for neuronal development in early preterm infants. The bi-state classification for AS and QS can be utilized for neural development indication and clinical decision support. As the minority states naturally occur less, they are of lesser importance to the course of development in the early stages of preterm infancy. In term infants, wake versus sleep becomes more important, but at that point, wake also has a more significant presence which can be utilized for training.

Compared to the current state of the art by Ansari et al., our method still lags in performance for QS and nonQS (AS) classification. However, it has to be kept in mind that AUC-ROC performance measure can be overestimating due to imbalanced data for majority and minority classes, which is incorporated in the here used kappa score. But again, a direct comparison is not in order as the methods are based on different physiological signals and follow a different goal. We believe that Ansari aims to classify QS and nonQS states in preterm infants with the highest possible performance and therefore chose the most promising and sleep-related physiologic signal to analyze. This manuscript, on the other hand, tries to classify preterm infant sleep states with a novel approach specifically based only on ECG derived features to ensure a more accessible and constant sleep monitoring. To emphasize our motivation again, the downside with using EEG for sleep analysis is the use of auxiliary sensors in addition to conventional monitoring sensors, which should be kept to a minimum as the preterm infant skin is highly sensitive. ECG, on the other hand, is a standard measurement in the NICU; therefore, this system is easily implementable in the current clinical practice. As the ECG signal is less interlocked with sleep compared to EEG, it is consequential that the results will fall behind, especially as this is the first ANN approach of this signal modality and patient group to date. Nevertheless, it is insightful to look at the publication from Ansari et al. as they also used an ANN approach and outperformed all former attempts on EEG based QS and nonQS sleep state classifications. This indicates that ANN approaches for preterm infant sleep classification are capable of outperforming traditional signal analysis and standard machine learning algorithms in general. Translating this to ECG signal analysis, we believe that with further tuning and additional data, also ECG based analysis with ANNs can reach outstanding performance levels. To our knowledge, this is the first approach to use an ANN approach in combination with ECG derived features for preterm infant sleep analysis. With further research, it is reasonable to believe that ECG based sleep analysis will become on par with human annotators and maybe EEG analysis.

The overall low performance on all state classification has to be explained by two rationals — first, the general difficulty in separating human physiological events that often show only a nuance of difference between its states, especially in preterm infants. This manifests in the existence of the sleep state IS, which by definition is a mix of the main sleep states AS, QS, and wake. IS incorporates patterns of all those sleep states making it very difficult to pinpoint the beginning and the end of neighboring states. Over the course of development, IS occurs less frequent, leaving a clearer picture of sleep state boundaries. This can also be seen in the generally low interrater variability, showing that trained observers have difficulties in uniformly identifying the sleep states. Some annotators had years of experience, having seen plenty of training data. The other core problem in preterm infants is the immaturity of the autonomic nervous system. This immaturity of the regulatory mechanisms leads to instability in regulation and control of ex- and internal stimuli. In preterm infants, those instabilities result in common heart rate decelerations that are not connected with an autonomous response to such as sleep state changes [57]. Nevertheless, they can easily be misinterpreted as such. The combination of those often only nuanced differences between autonomous states and the general instability in preterm infants inducing non-state-related heart rate changes makes the classification of preterm infant sleep a difficult task.

The second rationale is the low amount of data, especially for the minority classes. Preterm infant sleep around 70% in 24 h [10]. Therefore, the wake state is naturally underrepresented, and caretaking also takes only a portion of the day. Interestingly, QS is as well underrepresented among the three datasets but shows enough difference to AS to be sufficiently distinguishable. Generally, CTW shows differences in the patterns to the QS and IS states resulting in heightened performance for QS - CTW and IS - CTW classification despite the lack of data. The activity in both AS and CTW, and thereby signal similarity, makes it harder to classify resulting in the lowest performance. Another influence could be wrong annotations, as during CTW the preterm infant moves similarly to AS. If the eyes are not open or caretaking cannot be directly observed in the video frame, CTW could be mistaken for AS. Furthermore, AS - IS is better separable than QS – IS, which could be due to the reduced breathing and movement during IS. This reduction results in similar patterns for IS and QS, making the correct classification more difficult. Same as before, another reason could be the manual annotation. As changes to the heart rate variability indicating a state change without visible clues like twitches, eye movements or rapidly changing breathing, IS could be easily mistaken for the onset or continuation of QS.

ECG-based sleep state classification is far less studied compared to EEG. Most of the cardiorespiratory based work in preterm infants considered AS/QS or wake/sleep states, as all state classification is a challenging matter. To our knowledge, all state classification has not been investigated regarding classification, despite in our group. Reliable all state classification in preterm infants have yet to be presented.

The tri-state classification is expected to show slightly lower performance than the bi-state counterparts. Here, the more difficult states, IS and CTW, reduce the combined performance. The slightly higher performance between AS - QS - IS despite IS being a more difficult state to differentiate has to be explained with a higher amount of training data. Despite the lack of data, AS - QS - CTW classification shows only slightly reduced performance as noise, instability, and increased movement dominate the ECG patterns and create a clearer differentiation.

The performances on all state classifications using the residual approach are underwhelming. Nevertheless, the use of a simpler, sequential model also did not generate reasonable results. Here again, the data to train on, especially for the minority classes, was considerably small with very early, unstable, and fragile patients. However, due to higher performance on the majority classes AS and QS, general different feature modalities between the single states, and the very high difference in the amount of data between majority and minority classes shows that the problematic performance is directly linked to the data amount and not fundamental problems with the used model architectures. This generally indicates again that the correct track is to utilize deep learning for preterm infant sleep classification as deep learning has mostly a higher performance potential than machine learning with increasing data size as explained before. Generally, all-state classification is not of main importance for early preterm infant development monitoring but is vital for a holistic view on the patient's sleep rhythm and possible predictions of sleep patterns.

To summarize, the separation of AS and QS show the general potential of using deep learning for sleep classification based only on ECG derived features as stated in the research hypothesis. Similar results for the main sleep states were achieved compared to classic machine learning approaches [17,25]. Nevertheless, for a complete picture and overall sleep monitoring a wider study is necessary to gain a stable model including training on extreme outliers.

4.4. Model architectures

Recently, GRU networks were found to have similar performance as LSTM networks. The GRU network uses less computational power than the LSTM network, as it generates fewer parameters. Nevertheless, both units perform almost equally, and one cannot be generally favored over the other. We tested architectures with both units and found that in our case both, LSTMs and GRUs layer use, performed equally. Due to lower calculation time, GRU layers were used further on.

The wide and deep residual model architectures show similar results (Table 5). The total amount of layers after the initiation block is similar with ten layers in the deep model and 12 for the wide model. Both architectures consider low and high complexity relations between the features and sleep states with increasing hidden units. In the ResNext model approach, the idea was also to introduce cardinality, an increase of parallel structures per residual block. At this point, we only used a cardinality of one as the model architecture could not be enhanced further due to overfitting spiraling out of hand. A deep model with 25 GRU layers after the initiation block was run with massive overfitting problems. Compared to a model with 4 GRU layers after the initiation block, the residual structured models showed weaker performance (Table 3). The overall disappointing results of the residual architectures show that probably they overreach with the complexity of the analysis on that task at hand. The complexity cannot be put to use as too few training examples for the more complex feature conjunctions are at hand. Additionally, the increased complexity tends to lead to overfitting due to training onto complex appearing noise structures. Regularization and dropout have to be set in place that can lead to a performance restriction. A solution to finding the right model architecture might be an evolutionary approach for architecture search. Generally, using this novel combination of a ResNet/ResNext architecture with GRU layers and corresponding activation is promising and might be wide-ranging and valuable in highly complex time series analysis such as adult sleep analysis, where vastly more data is available.

Transfer learning did not result in an acceptable performance and did not improve the performance as hoped. Probably this is connected with the fact that the saved weights from the pre-trained models were taken from a single fold. Even though they showed reasonable results (Table 6), they lacked generalization on the validation data. Secondly, the data was further reduced by splitting the data pool for pre-training and later transfer learning for bias control.

4.5. Strength and limitations

It is a challenge to gather sufficient data in very preterm infants in the high-risk NICU environment. With a mean age of 29 ± 4.6 weeks GA the study group is very realistic and generalizable for a NICU population. Human annotators performed the annotation of the dataset with a moderate interrater-variability. This may limit the performance of an automated system from the very beginning. On the other hand, the trained model incorporates the different experiences and knowledge of different annotators creating a more stable, integral, and reliable model again. Interesting would be a wide range of annotators and annotation styles in future research, backed with sufficient data, to incorporate the derivations of different annotation techniques in the model. Due to this limited amount of data, the ANN approach could not develop its full potential; nevertheless, compared to the general low interrater variability in this patient group the results are acceptable. As said, the general strength of this approach is the potential to enhance its performance with increasing data.

Further, we believe that we have not yet found the features which describe the preterm infant sleep states in full detail representing their complexity. As novel feature extraction methods, such as using a CNN based on spectrogram, where used successfully in different ECG based applications (e.g., atrial fibrillation [58,59] or arrhythmia detection [60]), those methods for a better representation of the ECG signals, and a better way of extracting richer features from the ECG signals merits investigation.

4.6. Future perspectives

As the main reason for the low performance of the all-state classification can be linked with the low amount of data, considering the vast difference in preterm infant stability and development, we suggest that more preterm infant data has to be gathered to surmount the threshold where data size becomes not the primary influence on performance. Following, the gross amount of needed data is estimated. The assumption is that the classification performance would be similar for the data poorest class if such a state would have the same amount of data as now the data richest state. With wake as the data poorest state having 6.6% (not considering caretaking as it results from external influence) and AS as the data richest state now with 52.41%, the needed amount would be seven times higher as here present. This results in roughly 200 preterm infants with the same mean recorded time of 4 h. Alternatively, 50 patients with 24 h recordings, which would be more optimal regarding full sleep cycle analysis. As ANN performance is not linked linear to the data amount of a single state, it can be assumed that less data is sufficient.

Even though the transfer learning approach did not show the intended results, we suggest that term infant data instead of rare preterm infant data is used for pre-training as signal patterns and sleep architecture are still very similar to preterm infants and much more data is available for this patient group.

Another approach could be to look at unsupervised learning for preterm infant sleep staging. So far, we rely on human annotations, which are in itself not perfect and show large interrater variability. The general shift of data patterns from unsupervised learning could indicate brain development in the same way as classified state distributions from supervised learning. Unsupervised learning would demand even more data but will reduce the necessity of manual annotation. Not annotated, preterm infant sleep data is already freely available for example from the CHIME study [61].

5. Conclusions

Active and quiet sleep can be moderately separated using a deep learning approach solely using ECG derived features. Nevertheless, all state classification is, so far, not possible and is hindered mostly by limited preterm infant training data as well as training data of very young and unstable patients. There is a level of data that has to be reached so that the data amount is not the significant factor for performance. For highly complex time series analysis, backed up with sufficient data, an RNN-ResNet architecture with sigmoid activation can be chosen for a deep network approach, avoiding the problem of shattered gradients.

Declaration of Competing Interest

There are no competing interests to declare.

References

[1] J. Werth, L. Atallah, P. Andriessen, X. Long, E. Zwartkruis-Pelgrim, R.M. Aarts, Unobtrusive sleep state measurements in preterm infants – a review, Sleep Med. Rev. 32 (April) (2017) 109–122.

[2] D. Jouvet-Mounier, L. Astic, D. Lacote, Ontogenesis of the states of sleep in rat, cat, and guinea pig during the first postnatal month, Dev. Psychobiol. 2 (January (4)) (1970) 216–239.

[3] P. Peirano, C. Algarín, R. Uauy, Sleep-wake states and their regulatory mechanisms throughout early human development, J. Pediatr. 143 (October (4) Suppl) (2003) S70–9.

[4] M. Mirmiran, E. Van Someren, The importance of REM sleep for brain maturation, J. Sleep Res. 2 (1993) 188–192.

[5] V. Doria, C.F. Beckmann, T. Arichi, N. Merchant, M. Groppo, F.E. Turkheimer, S.J. Counsell, M. Murgasova, P. Aljabar, R.G. Nunes, D.J. Larkman, G. Rees, D. Edwards, Emergence of resting state networks in the preterm human brain, Proc. Natl. Acad. Sci. 107 (November (46)) (2010) 20015–20020.

[6] G. Calciolari, R. Montirosso, The sleep protection in the preterm infants, J. Matern. Fetal. Neonatal. Med. 24 (October (Suppl. 1)) (2011) 12–14.

[7] S. Graven, Sleep and brain development, Clin. Perinatol. 33 (September (3)) (2006) 693–706.

[8] M.S. Scher, Ontogeny of EEG-sleep from neonatal through infancy periods, Sleep Med. 9 (August (6)) (2008) 615–636.

[9] L. Curzi-Dascalova, P. Peirano, F. Morel-Kahn, Development of sleep states in normal premature and full-term newborns, Dev. Psychobiol. 21 (July (5)) (1988) 431–444.

[10] A.W. De Weerd, R.A.S. Van den Bossche, The development of sleep during the first months of life, Sleep Med. Rev. 7 (April (2)) (2003) 179–191.

[11] D. Holditch-Davis, M.S. Scher, T. Schwartz, D. Hudson-Barr, Sleeping and waking state development in preterm infants, Early Hum. Dev. 80 (October (1)) (2004) 43–64.

[12] C.L. Booth, H.L. Leonard, E.B. Thoman, Sleep states and behavior patterns in preterm and fullterm infants, Neuropediatrics 11 (November (4)) (1980) 354–364.

[13] A. Ansari, O. De Wel, M. Lavanga, A. Caicedo, A. Dereymaeker, K. Jansen, J. Vervisch, M. De Vos, G. Naulaers, S. Van Huffel, Quiet sleep detection in preterm infants using deep convolutional neural networks, J. Neural Eng. 15 (6) (2018).

[14] N. Koolen, L. Oberdorfer, Z. Rona, V. Giordano, T. Werther, K. Klebermass-Schrehof, N. Stevenson, S. Vanhatalo, Automated classification of neonatal sleep states using EEG, Clin. Neurophysiol. 128 (June (6)) (2017) 1100–1108.

[15] A. Gruetzmann, S. Hansen, J. Müller, Novel dry electrodes for ECG monitoring, Physiol. Meas. 28 (November (11)) (2007) 1375–1390.

[16] L. Atallah, A. Serteyn, M. Meftah, M. Schellekens, R. Vullings, J.W.M. Bergmans, A. Osagiator, S.B. Oetomo, Unobtrusive ECG monitoring in the NICU using a capacitive sensing array, Physiol. Meas. 35 (May (5)) (2014) 895–913.

[17] J. Werth, A. Serteyn, P. Andriessen, R.M. Aarts, X. Long, Automated preterm infant sleep staging using capacitive electrocardiography, Physiol. Meas. (2019).

[18] P. Fonseca, X. Long, M. Radha, R. Haakma, R.M. Aarts, J. Rolink, Sleep stage classification with ECG and respiratory effort, Physiol. Meas. 36 (October (10)) (2015) 2027–2040.

[19] P. Fonseca, N. den Teuling, X. Long, R.M. Aarts, Cardiorespiratory sleep stage detection using conditional random fields, IEEE J. Biomed. Heal. Informatics 21 (July (4)) (2017) 956–966.

[20] M. Radha, G. Garcia-Molina, M. Poel, G. Tononi, Comparison of feature and classifier algorithms for online automatic sleep staging based on a single EEG signal, 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2014) 1876–1880.

[21] X. Long, P. Fonseca, R.M. Aarts, R. Haakma, J. Rolink, S. Leonhardt, Detection of nocturnal slow wave sleep based on cardiorespiratory activity in healthy adults, IEEE J. Biomed. Heal. Informatics 21 (January (1)) (2017) 123–133.

[22] A. Procházka, J. Kuchyňka, O. Vyšata, P. Cejnar, M. Vališ, V. Mařík, Multi-class sleep stage analysis and adaptive pattern recognition, Appl. Sci. 8 (May (5)) (2018) 697.

[23] X. Long, P. Fonseca, R. Haakma, R.M. Aarts, Actigraphy-based sleep/wake detection for insomniacs, 2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN) (2017) 1–4.

[24] A. Lewicke, E. Sazonov, M.J. Corwin, M. Neuman, S. Schuckers, Sleep versus wake classification from heart rate variability using computational intelligence: consideration of rejection in classification models, IEEE Trans. Biomed. Eng. 55 (1) (2008) 108–118.

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

[26] J. Schmidhuber, Deep Learning in neural networks: an overview, Neural Netw. (2015).

[27] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (November (8)) (1997) 1735–1780.

[28] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, Circuits Syst. (2006), APCCAS 2006. IEEE Asia Pacific Conf., Jun. 2014.

[29] S. Biswal, J. Kulas, H. Sun, B. Goparaju, M.B. Westover, M.T. Bianchi, J. Sun, SLEEPNET: automated sleep staging system via deep learning, Comput. Res. Repository (2017), vol. eprint. 25-Jul-2017.

[30] S. Chambon, M.N. Galtier, P.J. Arnal, G. Wainrib, A. Gramfort, A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series, IEEE Trans. Neural Syst. Rehabil. Eng. 26 (April (4)) (2018) 758–769.

[31] A. Supratak, H. Dong, C. Wu, Y. Guo, DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG, IEEE Trans. Neural Syst. Rehabil. Eng. 25 (November (11)) (2017) 1998–2008.

[32] A.N. Olesen, P.E. Peppard, H.B. Sorensen, P.J. Jennum, E. Mignot, End-to-End deep learning model for automatic sleep staging using raw PSG waveforms, Sleep 41 (1) (2018) A121.

[33] W. Chen, A. Sano, D.L. Martinez, S. Taylor, A.W. McHill, A.J.K. Phillips, L. Barger, E.B. Klerman, R.W. Picard, Multimodal ambulatory sleep detection 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), vol. 40, 2017, pp. 465–468.

[34] M. Radha, P. Fonseca, M. Ross, A. Cerny, P. Anderer, R.M. Aarts, LSTM knowledge transfer for HRV-based sleep staging, Quant. Biol. (September (9)) (2018) 1–11, vol. eprint.

[35] H.F.R. Prechtl, The behavioural states of the newborn infant (a review), Brain Res. 76 (10) (1974) 185–212.

[36] R. Wijshoff, M. Mischi, R.M. Aarts, Reduction of periodic motion artifacts in photoplethysmography, IEEE Trans. Biomed. Eng. 64 (1) (2017) 196–207.

[37] T. Ruf, The Lomb-Scargle periodogram in biological rhythm research: analysis of incomplete and unequally spaced time-series, Biol. Rhythm Res. 30 (April (2)) (1999) 178–201.

[38] P. Indic, E. Bloch-Salisbury, F. Bednarek, E. Brown, D. Paydarfar, R. Barbieri, Assessment of cardio-respiratory interactions in preterm infants by bivariate autoregressive modeling and surrogate data analysis, Early Hum. Dev. 87 (7) (2011) 477–487.

[39] S. Reulecke, Autonomic regulation during quiet and active sleep states in very preterm neonates, Front. Physiol. 3 (4) (2012) 1–9.

[40] R. Joshi, D. Kommers, X. Long, L. Feijs, S. Van Huffel, C. van Pul, P. Andriessen, Cardiorespiratory coupling in preterm infants, J. Appl. Physiol. (November) (2018), http://dx.doi.org/10.1152/japplphysiol.00722.2018.

[41] K. Cross, T. Oppe, The respiratory rate and volume in the premature infant, J. Physiol. 116 (February (2)) (1952) 168–174.

[42] F. Chollet, et al., Keras, 2015.

[43] A. Martin, A. Ashish, B. Paul, B. Eugene, C. Zhifeng, C. Craig, C. Greg S, D. Andy, D. Jeffrey, D. Matthieu, G. Sanjay, G. Ian, H. Andrew, I. Geoffrey, I. Michael, J. Yangqing, J. Rafal, K. Lukasz, K. Manjunath, L. Josh, M. Dan, M. Rajat, M. Sherry, M. Derek, O. Chris, S. Mike, S. Jonathon, S. Benoit, S. Ilya, T. Kunal, T. Paul, V. Vincent, V. Vijay, V. Fernanda, V. Oriol, W. Pete, W. Martin, W. Martin, Y. Yuan, Z. Xiaoqiang, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015.

[44] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Front. Psychol. 4 (December) (2015).

[45] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 1, 2017, pp. 5987–5995.

[46] D. Balduzzi, M. Frean, L. Leary, J. Lewis, K.W.-D. Ma, B. McWilliams, The shattered gradients problem: if resnets are the answer, then what is the question? Comput. Res. Repos. (February) (2017), vol. eprint.

[47] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.

[48] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, Comput. Res. Repos. (2014), vol.eprint.

[49] M.L. McHugh, Interrater reliability: the kappa statistic, Biochem. Med. 22 (3) (2012) 276–282.

[50] R. Joshi, B.L. Bierling, X. Long, J. Weijers, L. Feijs, C. Van Pul, P. Andriessen, A ballistographic approach for continuous and non-obtrusive monitoring of movement in neonates, IEEE J. Transl. Eng. Heal. Med. 6 (2018) 1–10.

[51] A. Kahn, B. Dan, J. Groswasser, P. Franco, M. Sottiaux, Normal sleep architecture in infants and children, J. Clin. Neurophysiol. 13 (3) (1996) 184–197.

[52] D.H. Crowell, L.J. Brooks, T. Colton, M.J. Corwin, T.T. Hoppenbrouwers, C.E. Hunt, L.E. Kapuniai, G. Lister, M.R. Neuman, M. Peucker, Infant polysomnography: reliability. Collaborative home infant monitoring evaluation (CHIME) steering committee, Sleep (1997).

[53] H. Danker-Hopfe, P. Anderer, J. Zeitlhofer, M. Boeck, H. Dorn, G. Gruber, E. Heller, E. Loretz, D. Moser, S. Parapatics, B. Saletu, A. Schmidt, G. Dorffner, Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard, J. Sleep Res. 18 (March (1)) (2009) 74–84.

[54] A. Dereymaeker, K. Pillay, J. Vervisch, S. Van Huffel, G. Naulaers, K. Jansen, M. De Vos, An automated quiet sleep detection approach in preterm infants as a gateway to assess brain maturation, Int. J. Neural Syst. 27 (September (6)) (2017), 1750023.

[55] J.R. Isler, T. Thai, M.M. Myers, W.P. Fifer, An automated method for coding sleep states in human infants based on respiratory rate variability, Dev. Psychobiol. 58 (2016) 1108–1115.

[56] L. Fraiwan, K. Lweesy, N. Khasawneh, M. Fraiwan, H. Wenz, H. Dickhaus, Time frequency analysis for automated sleep stage identification in fullterm and preterm neonates, J. Med. Syst. 35 (August (4)) (2011) 693–702.

[57] D. Kommers, R. Joshi, C. van Pul, L. Atallah, L. Feijs, G. Oei, S. Bambang Oetomo, P. Andriessen, Features of heart rate variability capture regulatory changes during kangaroo care in preterm infants, J. Pediatr. 182 (March) (2017).

[58] M. Zihlmann, D. Perekrestenko, M. Tschannen, Convolutional recurrent neural networks for electrocardiogram classification, Comput. Cardiol. (2017).

[59] J. Zhang, J. Tian, Y. Cao, Y. Yang, X. Xu, C. Wen, Fine-grained ECG classification based on deep CNN and online decision fusion, Comput. Res. Repos. abs/1901.0 (January) (2019).

[60] J.T. Ruiz, J.D.B. Pérez, J.R.B. Blázquez, Arrhythmia detection using convolutional neural models, in: Distributed Computing and Artificial Intelligence, 16th ed., 2019, pp. 120–127.

[61] Collaborative Home Infant Monitoring Evaluation, 1998 [Online]. Available: http://slone-web2.bu.edu/ChimeNisp/.