SLEEP, 2022, 1–11

https://doi.org/10.1093/sleep/zsac143
Advance Access Publication Date: 24 June 2022
Original Article

ORIGINAL ARTICLE

The Sleep Well Baby project: an automated real-time sleep–wake state prediction algorithm in preterm infants

Thom Sentner[1,†], Xiaowan Wang[2,†,•], Eline R. de Groot[2], Lieke van Schaijk[1], Maria Luisa Tataranno[2,3], Daniel C. Vijlbrief[2], Manon J.N.L. Benders[2,3], Richard Bartels[1,‡] and Jeroen Dudink[2,3,‡,*]

[1]Digital Health, University Medical Center Utrecht, Utrecht, The Netherlands, [2]Department of Neonatology, Wilhelmina Children's Hospital, University Medical Center Utrecht, Utrecht, The Netherlands and [3]Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands

[†]These authors contributed equally and are co-first authors.

[‡]These authors contributed equally and are co-corresponding authors.

[*]Corresponding authors. Jeroen Dudink, Department of Neonatology, Wilhelmina Children's Hospital, University Medical Center Utrecht, Utrecht, The Netherlands. Email: j.dudink@umcutrecht.nl; Richard Bartels, Digital Health, University Medical Center Utrecht, Utrecht, The Netherlands. Email: R.T.Bartels-6@umcutrecht.nl.

Abstract

Study Objectives: Sleep is an important driver of early brain development. However, sleep is often disturbed in preterm infants admitted to the neonatal intensive care unit (NICU). We aimed to develop an automated algorithm based on routinely measured vital parameters to classify sleep–wake states of preterm infants in real-time at the bedside.

Methods: In this study, sleep–wake state observations were obtained in 1-minute epochs using a behavioral scale developed in-house while vital signs were recorded simultaneously. Three types of vital parameter data, namely, heart rate, respiratory rate, and oxygen saturation, were collected at a low-frequency sampling rate of 0.4 Hz. A supervised machine learning workflow was used to train a classifier to predict sleep–wake states. Independent training (n = 37) and validation datasets were validation n = 9) datasets were used. Finally, a setup was designed for real-time implementation at the bedside.

Results: The macro-averaged area-under-the-receiver-operator-characteristic (AUROC) of the automated sleep staging algorithm ranged between 0.69 and 0.82 for the training data, and 0.61 and 0.78 for the validation data. The algorithm provided the most accurate prediction for wake states (AUROC = 0.80). These findings were well validated on an independent sample (AUROC = 0.77).

Conclusions: With this study, to the best of our knowledge, a reliable, nonobtrusive, and real-time sleep staging algorithm was developed for the first time for preterm infants. Deploying this algorithm in the NICU environment may assist and adapt bedside clinical work based on infants' sleep–wake states, potentially promoting the early brain development and well-being of preterm infants.

Statement of Significance

To the best of our knowledge, this is first time that a nonobtrusive, real-time bedside automated sleep–wake state classification method is developed for preterm infants. Compared to gold standard polysomnography, our algorithm doesn't require any additional electrodes and opens up unique opportunities for routine sleep assessment and long-term monitoring of sleep patterns in the NICU environment, which are especially valuable in improving the quality of neonatal health care and in facilitating neuroprotective Interventions. By tracking sleep-wake transitions in the real clinical setting, clinicians and caregivers are able to plan elective care based on the infant's sleep–wake states, thus protecting their sleep quality and promoting optimal early brain development.

Key words: preterm; sleep; machine learning; automated sleep staging; neonatal intensive care

Submitted: 16 February, 2022; **Revised:** 31 May, 2022

Introduction

Annually, 15 million infants are born before 37 completed weeks of gestational age (GA) [1]. Preterm infants experience a sudden environmental change, from the safe womb to an incubator in the neonatal intensive care unit (NICU), affecting critical developmental processes and may lead to lifelong problems [2]. During their NICU stay, preterm infants are sensitive to well-known risk factors for atypical neurodevelopment, such as stress and pain [3].

Sleep is essential for promoting optimal early brain development [4–6]. Four major sleep–wake states can be discerned in preterm infants: active sleep (AS), quiet sleep (QS), intermediate sleep (IS), and wake (W) [7, 8]. Given the different roles of each sleep–wake state in early development, it is necessary that the infant experiences all these states with sufficient quality and quantity [9, 10]. However, preterm infants in the NICU are usually exposed to a myriad of extrinsic stimuli that radically alter their sleep-wake states [11, 12]. Therefore, it would be ideal to schedule elective interventions and care procedures in the NICU based on the infants' sleep–wake states [13, 14].

Different sleep–wake states have unique behavioral and physiological characteristics [15]. AS is characterized by high motor activity levels, rapid brain activity, and irregular cardiorespiratory activity. The untrained observer can easily confuse AS with a W state. QS is characterized by low motor activity, slower brain activity, and more regular cardiorespiratory features. The W state can be subdivided into multiple stages or behaviors, such as drowsy, or crying.

Sleep–wake states in preterm infants are usually evaluated manually at the NICU, either by the direct observation of behavioral signs, or by the visual assessment of multiple physiological parameters recorded via polysomnography (PSG) [15–17]. The most valuable PSG parameters for preterm sleep assessment include electroencephalogram (EEG), electrocardiogram (ECG), and respiratory signals [16]. However, behavioral sleep observation and PSG methods are significantly limited by the time-consuming nature of training clinical staff and of the scoring process [17]. Moreover, the two manual sleep assessment techniques require significant expertise for final data interpretation. Thus, real-time sleep annotation results are scarce.

For these reasons, it is desirable to design a nonobtrusive, real-time bedside automated sleep–wake state classification method.

Machine learning (ML) classification techniques have exhibited great potential for identifying sleep–wake states automatically in preterm infants [18–24]. However, a common factor of the existing automated sleep staging methods is that they could not be easily implemented in NICUs, with the main concern being the availability of data. Most of these methods were developed based on multichannel EEG [18–23], which is not commonly used in NICUs and requires additional electrodes that may damage the vulnerable skin of preterm infants. Furthermore, these EEG-based algorithms were only able to distinguish between QS and non-QS stages despite the importance of AS in brain development [10, 25, 26].

Our group has demonstrated that cardiorespiratory characteristics, especially heart rate (HR) and respiration frequency, play an important role in distinguishing between sleep–wake states [27]. Until now, however, a reliable automated sleep staging model based on these parameters has not been readily available for preterm infants admitted to the NICU. A recent study conducted by Werth et al. [24] employed a convolutional neural network (CNN)-based algorithm combining HR variability (HRV), ECG data, and patient information to distinguish between sleep–wake states of preterm infants, with a kappa range between 0.33 and 0.44. However, using HRV and other ECG features requires data with a high sampling rate (250–500 Hz), which is not available at most NICUs [28].

The current study aimed to develop an automatic system with which preterm sleep–wake states can be easily classified at the bedside in NICUs. To broaden its applicability, our goal was to build a system based on vital physiological parameters at a low sampling rate, which is widely available in most NICUs.

Methods

Patient inclusion. The current study employed two independent datasets of preterm infants admitted to the NICU of the Wilhelmina Children's Hospital (WKZ), Utrecht, the Netherlands. The first dataset (37 infants; postmenstrual age [PMA] 31.1 ± 1.5 weeks) was used to train and evaluate the performance of the automated sleep classification algorithm, while the second dataset (9 infants; PMA 30.9 ± 1.3) was used to validate the algorithm. Exclusion criteria were: congenital malformations, seizures, major brain damage or abnormalities, and mother's use of recreational drugs during pregnancy. Written informed consent was obtained from parents before enrollment. Permission to use patient data was obtained from the Medical Ethical Review Committee of the University Medical Center (UMC) Utrecht (No. 21-066-C). Clinicians pseudonymized all data prior to the analysis.

Patient demographics were compared between the two datasets using Student's *t*-test (*t* value) for continuous variables, Mann–Whitney *U* test (*Z* score) for ordinal variables, and chi-squared test (χ^2) for categorical variables.

Sleep observation. A behavioral sleep–wake state classification system developed in-house was used for sleep observation, producing four different sleep–wake states: AS, QS, IS, and W [29]. Four observers performed sleep–wake state annotations for the training dataset and two other observers for the validation dataset. A single observer was present in each run. The observers recorded sleep–wake states at the end of a 1-minute time window. During observation, a confidence score of +1, 0, or –1 was assigned to each window, corresponding to high-to-low confidence level of an observer for a specific window. For both datasets, each infant was observed for approximately 3 consecutive hours. For the validation dataset, one infant was observed at two separate instances days apart.

Sleep–wake state and confidence score distributions were compared between the two datasets using multivariate analysis of variance (MANOVA) according to Wilks' lambda (Λ) statistic. Significant MANOVA tests were followed by univariate analysis of variance post-hoc tests, with *p*-values corrected using Bonferroni correction. The 95% confidence intervals (CIs) on the sleep–wake state distributions were generated from 250-fold bootstrapping with replacement. To better account for correlations between consecutive 1-minute observations, the full data from single observation runs of patients were resampled to create simulated populations rather than resampling individual 1-minute observations.

Physiological data collection. The vital physiological parameters were measured using IntelliVue MP70 patient monitors (Philips Healthcare, Best, the Netherlands) and recorded by a software solution developed in-house (BedBase, UMC Utrecht, the Netherlands). For each patient admitted to the NICU, HR, respiration rate (RR), and oxygen saturation (SpO_2) were routinely monitored.

Except for the HR in the validation dataset, all signals were acquired with a sampling rate of 0.4 Hz. Due to a software update, the sampling frequency of the HR for validation patients was 1 Hz. A down-sampling procedure from 1 Hz to 0.4 Hz was applied to these data to ensure a homogeneous database. For each included patient, 24-hour reference data were collected to correct signals for inter-patient variability. Given the rapid neonatal developmental changes, a 24-hour period with consecutive data availability was selected as close as possible in time to the observation period.

Data cleaning: observations. To avoid bias, all IS states (considered as an inherently noisy label), observations explicitly indicated as uncertain (i.e. confidence score of –1), and missing periods (e.g. due to a clinical intervention during the observation period) were excluded from further analysis. In total, 34.3% of raw observation data (2956/8605 min) were excluded.

Data cleaning: physiological parameters. Missing data points in the physiological data were removed from further analysis. Using the mean and standard deviation of the 24-hour reference data, HR and RR data were corrected for inter-patient variability by applying a standard scalar (Equation S1). Patients who did not have a consecutive 24-hour reference data in the 3.5 days before or after the observation period were excluded (7/37 patients in the training dataset; no patients in the validation dataset). A rescaling correction was applied to the reference data to ensure similar distributions between the 3-hour observation and reference period (see Supplementary Section 1). The reference correction procedure was not applied for validation patients to ensure generalizability to a real clinical setting.

Feature calculation: physiological parameters. Subsequently, a comprehensive set of time-series features was calculated based on HR, RR, and SpO_2 parameters using *tsfresh* for 60-, 120-, 240-, and 480-s time windows preceding each epoch. Observations with > 50% missing data for any of the feature windows were removed from further analysis (86/3251 min in the training set). For each feature window, the following nine statistics were calculated: median, mean, variance, maximum, minimum, linear trend slope, linear trend intercept, linear trend p-value, and linear trend R-value ($9\ statistics \times 3\ parameters \times 4\ time\ windows = 108\ features$).

ML model development and bedside implementation

Three commonly used classification methods were chosen for model development: logistic regression, decision tree, and random forest. Logistic regression is a simple model that makes predictions based on a linear combination of input features. Decision trees produce prediction rules by recursively dividing data into subgroups that are homogeneous until a final prediction is reached. Random forest is an ensemble of decision trees each trained on a random subset of the data and the prediction is made by majority voting. As an ensemble learning method, the random forest generally outperforms the other two classifiers but is less interpretable [30].

The three classifiers were trained using nested cross-validation (nCV) on the training dataset (see Figure 1 for an overview of the whole workflow). In both the inner and outer cross-validation loop, a fourfold grouped, stratified split was applied, with grouping on a patient level and stratification on the amount of wake states. Within the inner loop, training data were randomly oversampled on a patient level to equalize the amount of data per patient. Hyperparameters were optimized in the inner loop using grid search (see Supplementary Table S1). The outer loop was used to estimate out-of-sample performance for each model. The macro-averaged area-under-the-receiver-operator-characteristic (AUROC) was maximized and used to assess model performance. This metric represents the predictive performance of the classifier over all possible threshold values for all classes and is insensitive to the class imbalance present in our dataset [31].

After selecting the best model via the nCV procedure, the inner loop was applied once more using all training data. We used an elbow method to fix hyperparameters and to choose a less overfit model at the expense of a slight decrease in model performance (see Supplementary Section 2).

Subsequently, a regular cross-validation procedure was applied to the model with fixed hyperparameters, and prediction probabilities were calibrated using isotonic regression in a separate cross-validation procedure. Model calibration was evaluated using the Brier score (BS) and Brier skill score (BSS). The BS ranges from 0 to 1, with 0 implying that forecasted probabilities match observed probabilities. The BSS indicates whether the calibration of one model is better than that of another, with higher values indicating better performance. Model predictions for training samples were calculated using cross-validation to better balance out-of-sample errors and ensure generalizability of results. Feature importance was assessed through SHapley Additive exPlanations (SHAP) values [32]. A final calibrated model was fitted on all training data and applied to the validation dataset. The 95% CIs on the model performance were generated from 250-fold bootstrapping with replacement of whole observation runs of patients.

Finally, as the aim of the current study was to develop a sleep–wake state classification algorithm that can be continuously available for preterm infants at the NICU, an architecture was designed to allow real-time predictions at the bedside. Data version control was applied to guarantee proper architecture functionality, and code tests were written.

Comparison with previous state-of-the-art methods. To provide a clear view of the pros and cons of the proposed sleep–wake state classification algorithm, we compared it with seven state-of-the-art methods that were previously developed for preterm sleep staging. Werth et al. [24] employed a sequential CNN model to classify AS and QS using ECG and HRV features. Koolen et al. [18] trained a support vector machine (SVM) classifier with a set of EEG characteristics as input to identify AS and QS. The remaining five studies were carried out by the same research group. They used eight-channel EEG signals to detect QS from non-QS states, in which least squares SVM (LS-SVM) [19],

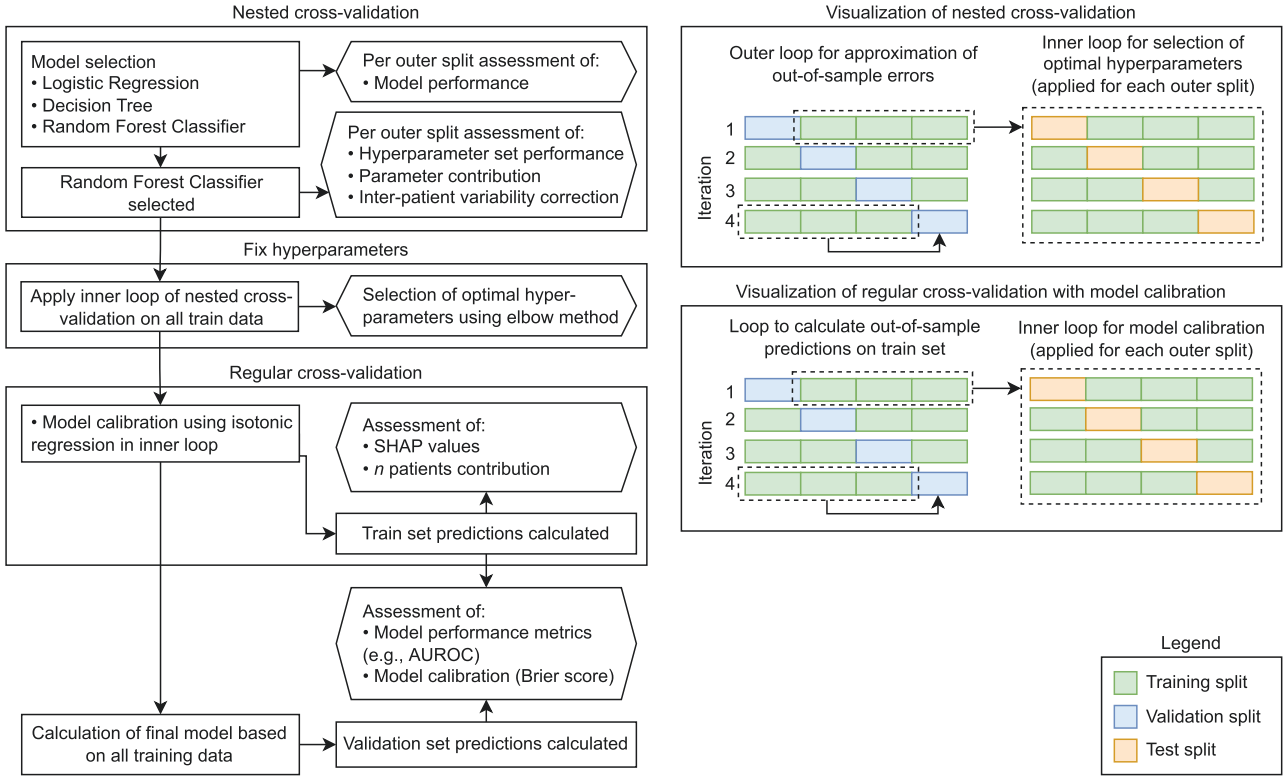

Figure 1. Schematic representation of the machine-learning model development. The nested cross-validation procedure was applied to select the random-forest classifier over other models and identify its optimal hyperparameters. The regular cross-validation procedure was applied to approximate out-of-sample performance using the training dataset, which was later validated using the validation dataset.

Table 1. Summarized patient demographics, observation, and parameter data, for the training and validation datasets.

Variable	Training patients (*N* = 30)	Validation patients (*N* = 9)
Gender, male, % (*n*)	60% (18)	56% (5)
Delivery method, % (*n*)		
Spontaneous	37% (11)	44% (4)
Cesarean section	47% (14)	33% (3)
Medicinally induced labor	0% (0)	11% (1)
Unknown	17% (5)	11% (1)
Birth weight, g, mean ± SD	1156 ± 368	1215 ± 355
Multiple birth, % (*n*)	33% (10)	0% (0)
Apgar score, median (Q_1, Q_3) [*n* unknown]		
At 1 min	5 (3, 7) [8]	4 (3, 6) [5]
At 5 min	7 (6, 8) [8]	6 (4, 6) [5]
At 10 min	8 (8, 9) [11]	8 (5, 8) [5]
Gestational age, weeks, mean ± SD	28.9 ± 2.0	29.0 ± 1.4
Postmenstrual age at study, weeks, mean ± SD	31.2 ± 1.5	30.9 ± 1.3
Observation period length per patient, minutes, median (Q_1, Q_3)	180 (178, 180)	180 (180, 180)
Observed sleep-wake states, % (*n*)		
Confidence score −1	19.5% (1035)	7.8% (138)
Confidence score 0	45.7% (2424)	73.5% (1307)
Confidence score 1	34.9% (1851)	18.7% (333)
Observed sleep-wake states with confidence score ≥0, % (X – X%, 95% C.I.) (*n*)		
Active sleep	37.1% (30.5–43.0%) (1584)	48.3% (43.2–54.1%) (792)
Intermediate sleep	21.9% (19.4–24.7%) (938)	23.0% (18.7–27.7%) (378)
Quiet sleep	32.7% (27.0–38.8%) (1399)	21.0% (16.8–25.7%) (344)
Wake	8.3% (5.3–12.1%) (354)	7.7% (3.4–11.9%) (126)

cluster-based adaptive sleep staging (CLASS) [20], deep CNN [21], end-to-end CNN [22], and multiscale deep CNN [23] techniques were adopted, respectively.

Software. Preprocessing and modeling were performed in Python 3.7 using the packages NumPy 1.20.3, pandas 1.2.5, scikit-learn 1.0.1, SciPy 1.6.3, shap 0.36.0, statsmodels 0.11.1, and tsfresh 0.17.0.

Results

Patient demographics. Patient characteristics are shown in Table 1, with respective statistical tests summarized in Supplementary Table S2. The training dataset consisted of 30 patients ranging between 28.3 and 33.5 weeks PMA. The validation dataset comprised 9 patients ranging between 29.0 and 33.3 weeks PMA. None

Table 2. Overview of performance metrics of the regular cross-validation procedure and validation dataset, with 95% CI values between brackets, calculated using 250-fold bootstrapping. Any metrics indicated with a sleep–wake state were calculated for that specific sleep–wake state vs the combination of other sleep–wake states.

	Training dataset	Validation dataset
Balanced accuracy Wake	70.0% (61.0–77.0%)	63.0% (47.0–77.0%)
Sensitivity Wake	47.0% (29.0–60.0%)	43.0% (8.0–74.0%)
Specificity Wake	93.0% (88.0–96.0%)	84.0% (69.0–93.0%)
F1 score (macroaveraged)	59.0% (50.0–66.0%)	48.0% (37.0–57.0%)
F1 score Wake	46.0% (26.0–61.0%)	29.0% (5.0–50.0%)
AUROC (macro-averaged)	76.0% (69.0–82.0%)	70.0% (61.0–78.0%)
AUROC Active Sleep	68.7% (59.1–75.9%)	66.0% (58.7–72.8%)
AUROC Quiet Sleep	78.3% (72.6–84.0%)	66.8% (54.2–78.1%)
AUROC Wake	80.0% (70.0–88.0%)	77.0% (63.0–87.0%)
Cohen's kappa	0.376 (0.259–0.472)	0.235 (0.089–0.357)
Cohen's kappa Active Sleep	0.314 (0.186–0.417)	0.232 (0.108–0.361)
Cohen's kappa Quiet Sleep	0.434 (0.325–0.528)	0.268 (0.110–0.414)
Cohen's kappa Wake	0.398 (0.189–0.553)	0.188 (–0.042–0.42)
Brier Active Sleep	0.224 (0.203–0.254)	0.243 (0.216–0.276)
Brier Quiet Sleep	0.185 (0.159–0.215)	0.190 (0.165–0.219)
Brier Wake	0.082 (0.064–0.103)	0.108 (0.074–0.147)

Table 3. Comparison of the proposed SWB algorithm to the state-of-the-art preterm sleep-prediction methods.

Algorithm		Data type	Sampling rate	AUC%	Mean kappa (SD)	State pairs	Pros	Cons
Vital signs-based classifier	SWB (random forest)	HR, RR, SpO2	0.4 Hz	76	0.38 (0.05)	AS–QS–W	Allow real-time and bedside monitoring in the NICU; allow AS and W prediction; only needs routine data with low sampling rate	Moderate sleep-prediction performance
	Sequential CNN [24]	HRV and other ECG-derived features	500 Hz	–	0.43 (0.08)	AS–QS	Relatively higher prediction performance; allow AS prediction	The required ECG signals with high sampling rate are expensive and not commonly available in NICUs; no W prediction
EEG-based classifier	SVM (feature-based) [18]	8-channel EEG	256 Hz	88		AS–QS	Relatively higher prediction performance; allow AS prediction	The required EEG signals are expensive and not commonly available in NICUs; additional sensors required; holds challenges for long-term routine monitoring; no W prediction
	LS-SVM (feature-based) [19]	8-channel EEG	250 Hz	90		QS–non-QS	Relatively higher prediction performance	The required EEG signals are expensive and not commonly available in NICUs; additional sensors required, holds challenges for long-term routine monitoring; no AS and W prediction
	CLASS (cluster-based) [20]			92	0.66 (0.24)			
	Deep CNN [21]			93	0.68 (0.22)			
	End-to-end CNN [22]			95	0.76 (0.22)			
	Sinc (multiscale deep CNN) [23]			88	0.77 (–)			

of the patient characteristics differed significantly (p-values ≥ .05) apart from the per-epoch confidence score counts (p-value < .001). Furthermore, W was the clear minority class, observed in 8.3% of time points in the training dataset, while AS and QS were observed 37.1% and 32.7%, respectively.

Model performance. The following macroaveraged AUROCs for the three classifiers were obtained from the nCV procedure: 0.710 ± 0.039 (decision tree), 0.750 ± 0.044 (logistic regression), and 0.782 ± 0.038 (random forest). Random forest was selected as the reference model, as it performed best, and its hyperparameters were fixed (see Supplementary Section 2). Subsequent discussion will refer to this model. Model generalization was tested by applying the random-forest model to the validation dataset. An extensive overview of performance metrics can be found in Table 2. The CIs of all metrics overlap between the two datasets, indicating that the model generalizes to unseen data. The properties and performance of our algorithm are compared with recently proposed state-of-the-art sleep staging methods in Table 3.

The macroaveraged AUROCs of the random-forest classifier are illustrated in Figure 2. Prediction performance was best for W in both datasets, with an AUROC of 0.80 (95% CI, 0.70–0.88) in the training dataset and 0.77 (95% CI, 0.63–0.87) in the validation dataset. Supplementary material, Section 4.1 discusses the confusion matrix, which further illustrates that it is unlikely that W gets confused with QS, and that AS is predicted with the highest precision, at 75%, when fixing the prediction threshold.

Figure 3 shows the calibration curves, that is, the correspondence between predicted and actual probabilities, for model predictions of the sleep–wake states. The predictions on the cross-validated training data appear well calibrated. Compared to a model without isotonic calibration, BS scores of 11.5% (AS), 6.5% (QS), and 25.8% (W) were observed, indicating that calibration was warranted. For a calibrated model predicted probabilities can be summed to estimate the total time spent in a sleep–wake state over extended time periods. It should be noted that the spread in accuracy among patients is still substantial (Figure 3, bottom panels). Furthermore, predictions for the validation dataset appear less well calibrated, which can, in part, be attributed to the small number of patients (see Supplementary material, Section 5).

To explore feature contributions, SHAP values are shown in Figure 4. HR-related features dominate the top 10 features for each sleep–wake state. A model based solely on HR-related

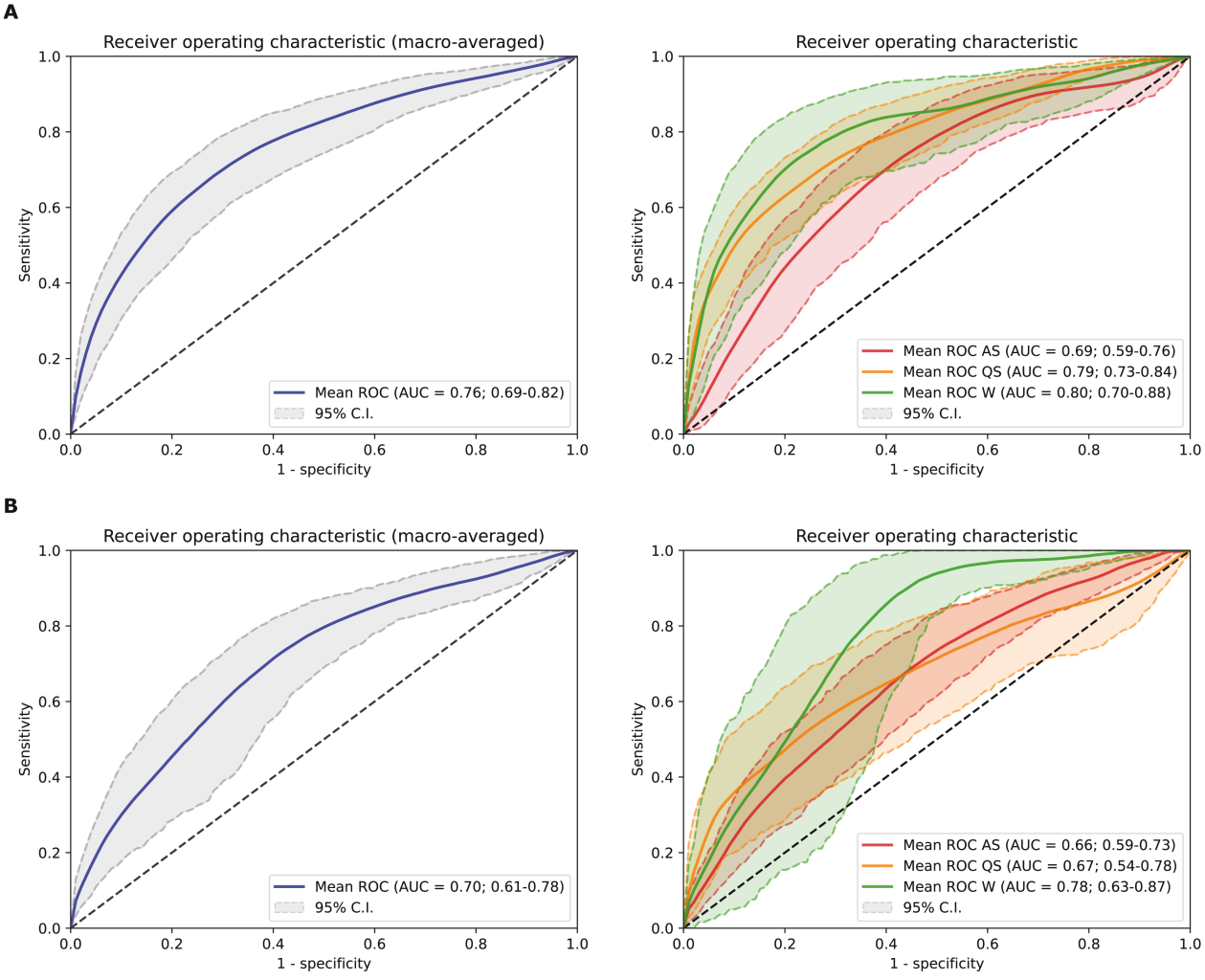

Figure 2. AUROCs calculated for the random forest classifier using (A) the regular cross-validation procedure on the training dataset and (B) the validation dataset. Performance mean and 95% Confidence Intervals (solid lines and filled areas, respectively) were calculated using 250-fold bootstrapping. Left, macroaveraged over all sleep–wake states; right, for the individual sleep–wake states.

features performs similar to a model including all parameters (see Supplementary material, Section 4.2). As expected, some features contribute oppositely to different sleep–wake states. For instance, a higher maximum HR during the last 480 s indicates a higher probability of W but a lower probability of QS. Moreover, while the top features of W are strongly dominated by amplitude-related HR features, variance of HR and SpO_2 contribute most to QS and AS predictions.

In Supplementary material, Sections 4.3–4.4, additional model performance checks are discussed.

Bedside implementation. Figure 5 illustrates a bedside implementation architecture currently being tested at the NICU of the WKZ, UMC Utrecht. Data are collected and stored in a database by the BedBase application. Every minute, BedBase sends a POST request to the sleep-prediction application: Sleep Well Baby (SWB). The POST request includes vital parameter data (HR, RR, SpO_2) and patient age encoded in JSON format. Inside the SWB application, several checks are performed. Validity checks are made for the vital parameters by ascertaining that they fall within the expected physical ranges and ensuring that the percentage of missing data is <50%, in line with the model design mentioned above. In addition, the PMA is used to check whether the patient falls within the age range of 28–34 weeks PMA, for which SWB was developed. If all criteria are met, SWB returns a prediction of "wake", "active sleep", or "quiet sleep". Additional eligibility criteria related to the patient's health—such as receiving invasive respiratory support—are to be verified by healthcare professionals. BedBase stores the predictions in the database and visualizes the current prediction and historical trends. Furthermore, BedBase returns probabilities per sleep–wake state.

Discussion

An automated, real-time sleep staging algorithm, called Sleep Well Baby, was developed for preterm infants. The algorithm was based on HR, RR, and SpO_2 signals sampled in low frequency and identified W, QS, and AS states. The random-forest classifier achieved good performance on the training dataset

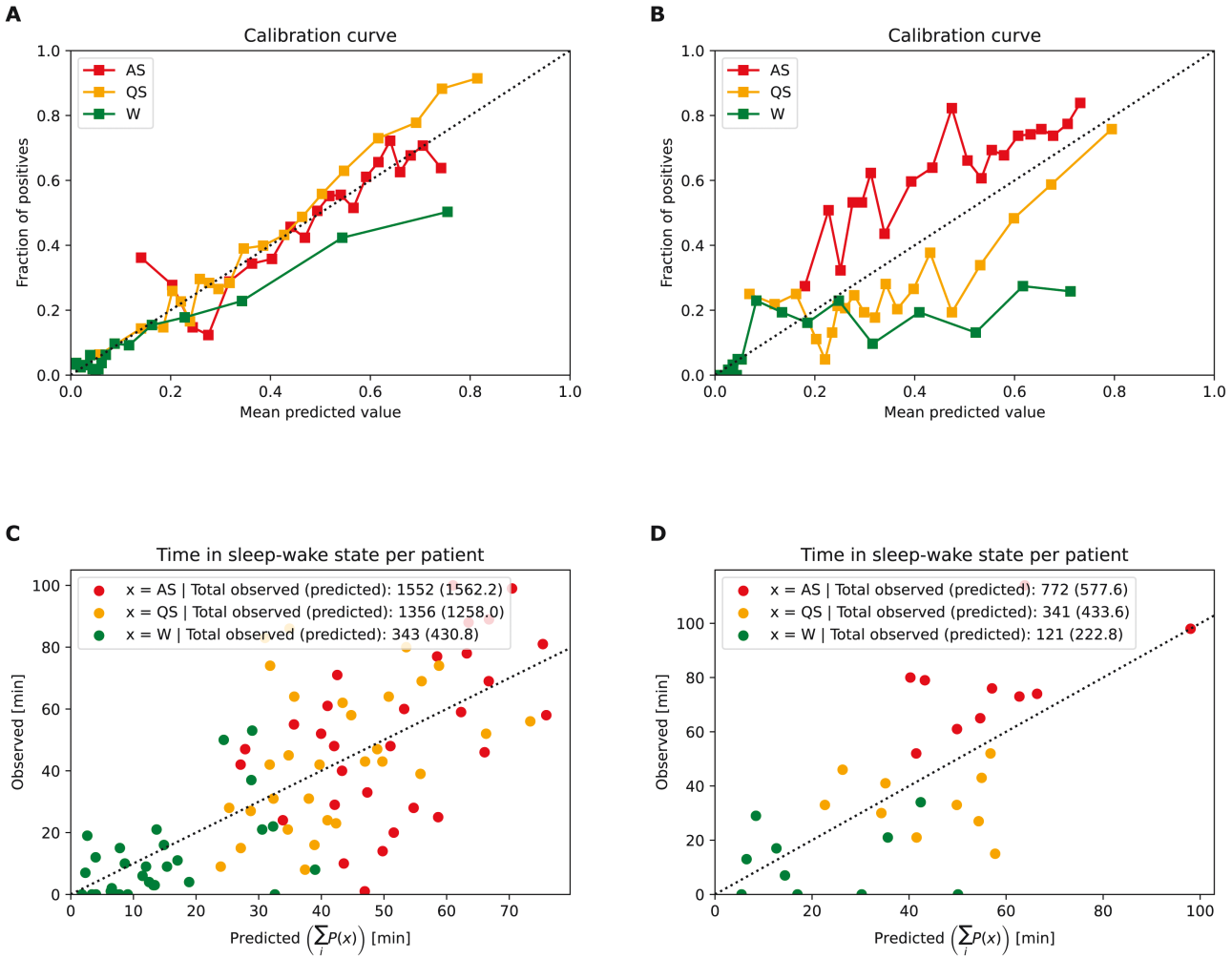

Figure 3. Top panels: model calibration per sleep–wake state (W = Wake, AS = Active Sleep, QS = Quiet Sleep), with data points split over 20 quantiles (bins) per sleep–wake state, for (A) training dataset and (B) validation dataset. The horizontal axis indicates the mean value of all predicted probability in each bin and the vertical axis indicates the proportion of positive class samples in each bin. Bottom panels: predicted time spent in a sleep–wake state (horizontal axis) versus the observed time in the sleep–wake state (vertical axis). Dots represent individual patients and are colored by sleep–wake state (Green/W = Wake, Orange/AS = Active Sleep, Red/QS = Quiet Sleep). Predicted time corresponds to summation of all model probability predictions. Panel (C) shows results for the training dataset. Panel (D) shows the results for the validation dataset.

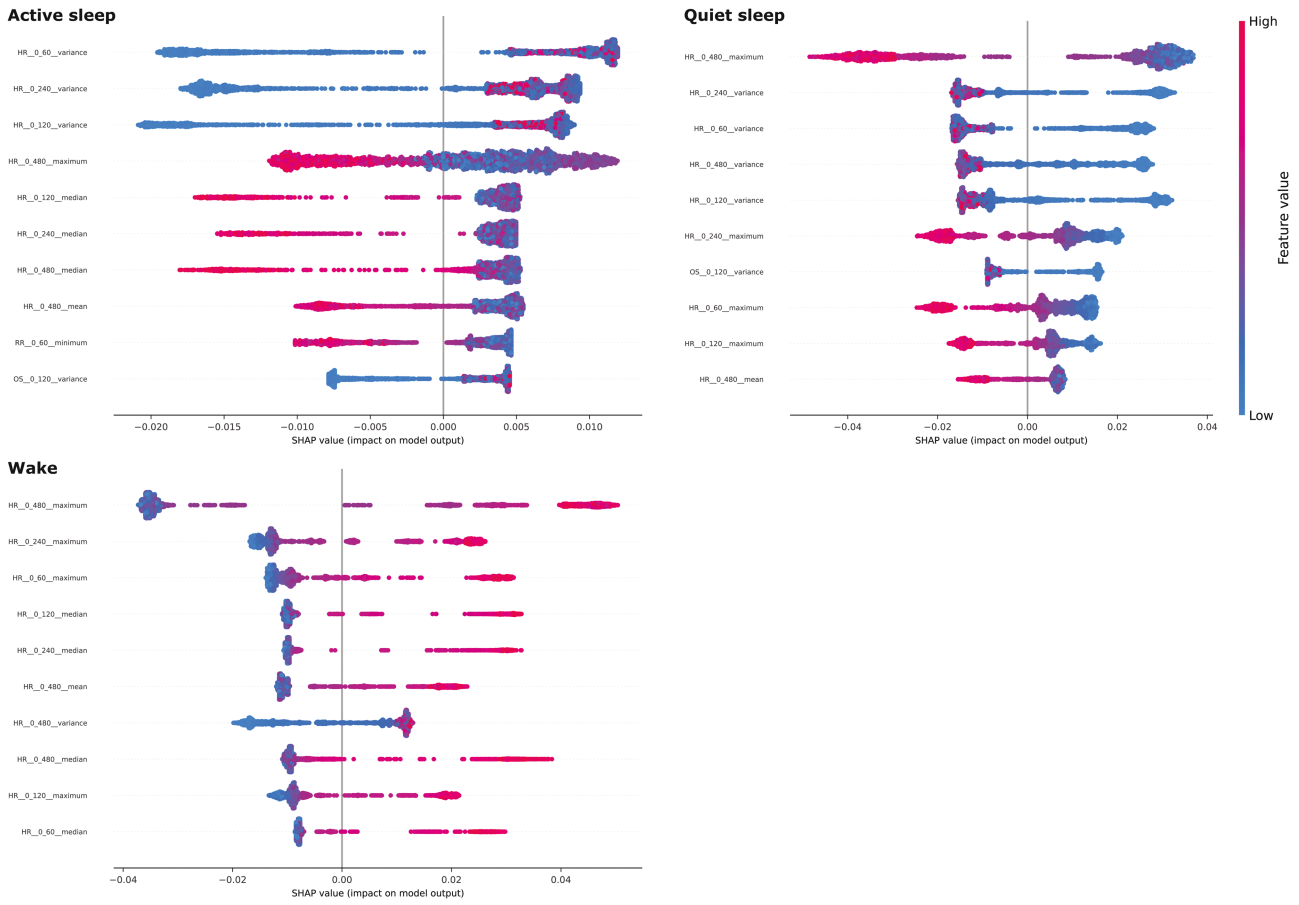

Figure 4. SHAP values of top-10 features of Active sleep (top left), Quiet sleep (top right) and Wake (bottom left). Naming convention is <vital parameter abbreviation>__<window start (seconds)>_<window end (seconds)>__<feature>. Dots represent individual data points. Feature values are indicated on a red-blue scale, with red representing a high value for that specific feature. Given the feature and its value for a specific data point, the impact on the model output (SHAP-value) is depicted on the *x*-axis.

(macroaveraged AUROC range 0.69–0.82), which was validated in an independent dataset (macro-averaged AUROC range 0.61–0.78).

To the best of our knowledge, the current algorithm is the first to combine multiple cardiorespiratory measures for automated preterm sleep–wake classification. The predictive performance of the SWB algorithm (training dataset $\kappa = 0.38 \pm 0.05$, validation dataset $\kappa = 0.24 \pm 0.07$) is comparable to previous work focusing on cardiorespiratory parameters. Research conducted by Werth et al. [24] showed that AS and QS could be distinguished based on a CNN algorithm using HRV, with a Cohen's kappa of $\kappa = 0.43 \pm 0.08$ (Table 3). Notably, the SWB algorithm was able to achieve similar performance to this method, but with less complex computations and without the need for high sampling frequency measurements that are expensive and not commonly available in NICUs. When comparing the SWB algorithm to existing sleep staging methods based on EEG [18, 21], there is still room for improvement in model performance (Table 3). Nonetheless, our algorithm may be applied with minimal intrusion and at most NICUs, as it uses routinely collected data without additional requirements.

Feature contribution analyses showed that HR contributes the most in the prediction of sleep–wake states. HR is regulated by the autonomic nervous system (ANS). As sleep–wake state changes, the ANS controls the heart directly, inducing immediate HR changes and affecting the respiratory and other physiological systems [33]. Our results, therefore, provided evidence that HR might play a more important role than other parameters in sleep–wake autonomic regulation of preterm infants. Oxygen saturation was also included in the algorithm, although it was not expected to be directly affected by the sleep–wake cycle based on literature. However, it did show some predictive value, possibly because of its close link with breathing.

Mispredictions of the SWB model often involve AS classifications when the infant is awake. Conversely, AS is predicted with the greatest precision of the sleep–wake states, at 75%. Since it is believed that AS is important for early brain maturation [10, 25, 26], it is of utmost priority to not disturb this sleep state during infants' stay in the NICU. In this context, mispredictions of our model tend to align positively with clinical impact. In addition, qualitative analysis of the time-ordered predictions for individual patients showed that a large percentage of mispredictions was only a few minutes ahead or behind the actual time of observation, for example, a wake prediction one minute too early (Supplementary material, Section 4.4). As such, the sleep–wake transition trend was still well captured by the SWB model, with little negative effect on its practical application. Consequently, despite the possible incorrect predictions made by the SWB model, it constitutes a helpful and practical implementation in the NICU.

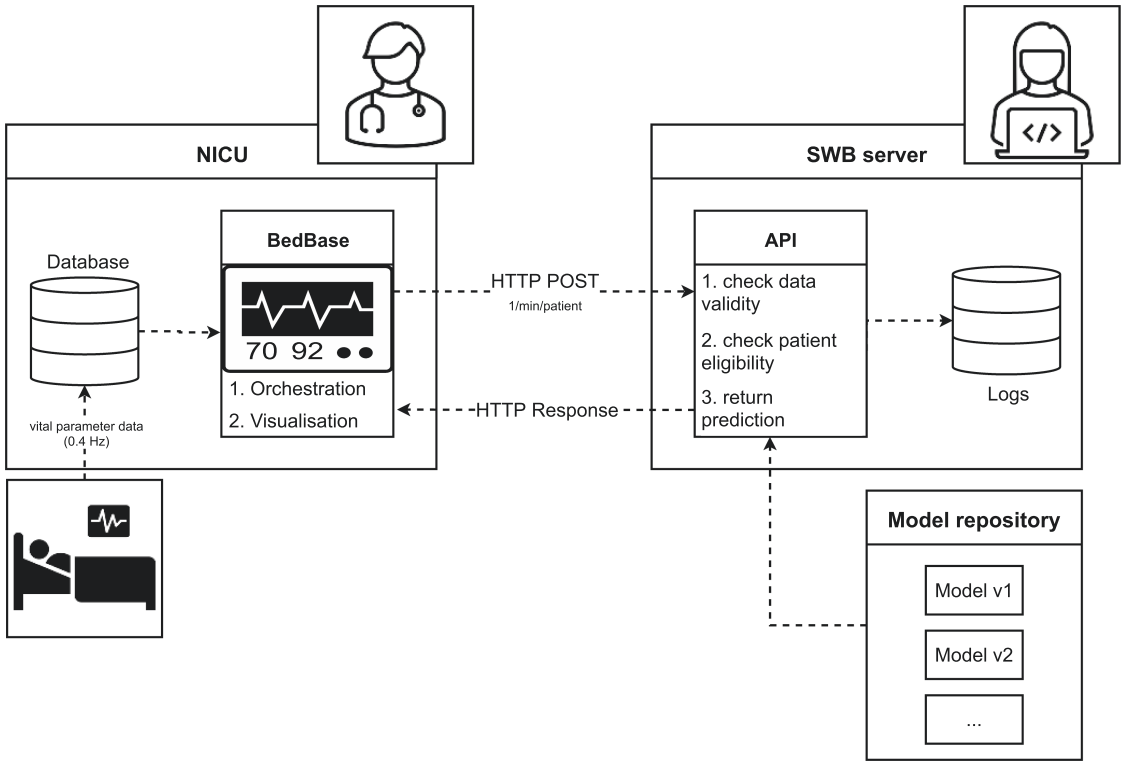

Figure 5. Bedside implementation architecture. Every minute, the BedBase application extracts real-time patient information and sends a POST request to the SWB API. Other than age, no identifiable data leaves the NICU. Before each prediction the SWB application assures that the data is valid (e.g. no more than 50% missing values per parameter in each time window, range of allowed values per parameter, etc.) and checks that the patient age falls within the range the algorithm was designed for (28–34 weeks postmenstrual age). If all criteria are met a sleep-state prediction is made and the result returned. Aggregated performance metrics are logged, e.g. ratio of wake predictions per day, such that an ICT and data scientists can monitor performance without compromising patient privacy. Data and code version control is implemented to ensure the SWB API always uses the correct model version from the model repository.

Several limitations need to be mentioned. Compared to statistical models usually used in medical research, such as linear regression, the random-forest model used in this study is less interpretable and explanations for its predictions require additional techniques such as SHAP. However, the SWB algorithm is designed to improve the quality of neonatal health care by continuously monitoring sleep–wake states, rather than making high-stake decisions. It does so by only presenting the predicted sleep–wake state, without further explanation on how it came to its conclusion. Given its intended use and since model predictions are presented independently without explanation, we believe it is reasonable to employ a black-box model providing better prediction performance over a more interpretable model. Another limitation is that the behavioral sleep observations were performed by one observer at a time in this study to prevent crowding in the NICU. Despite the overall high reliability and validity of the employed behavioral sleep–wake observation scale [29], this could inherently limit the potential performance of our model because some observers only achieved moderate reliability (κ = 0.46–0.82). Nonetheless, by including various experiences and knowledge of different observers, the SWB model may transcend the limitations of individual observers and be considered more stable and reliable [24]. Finally, to create better reference profiles, we only included those who had consecutive 24-hour vital signs data closely before or after the sleep observation, which resulted in the limited sample size. A larger patient cohort is beneficial to better assess and improve calibration performance. Although we do not expect improvements in discriminative power from a larger patient sample (see Supplementary material, Section 4.3), the isotonic calibration procedure benefits from a larger patient cohort [34]. Since isotonic calibration of the model can be performed after model training, future sleep–wake state observations of new patients can be used to improve calibration without affecting discriminative power (see Supplementary material, Section 5).

The SWB algorithm is being implemented and tested in our NICU at the bedside, with which nurses are able to plan elective care based on the sleep–wake state predictions and future research is able to further explore whether sleep protection in clinical practice improves preterm infants' outcomes. Moreover, cardiorespiratory parameters are a convenient starting point for automated sleep–wake state classification. It would be interesting for future research to combine them with other typical sleep-relevant features, such as body twitches unobtrusively captured by video analysis. In addition to real-time application in the NICU, the concept of assessing sleep stages based on cardiorespiratory parameters with low sampling rates also makes the SWB algorithm an attractive solution in a plethora of situations. For example, the algorithm can be applied in pre-clinical animal studies such as preterm lambs, or be adjusted for (preterm) infants once they are at home.

In summary, the current study developed an automated, real-time sleep staging system, called SWB, for preterm infants based on low-frequency vital physiological parameters, which are most often monitored in the NICU. A three-class classifier was designed to predict the infant's sleep–wake states. The SWB

system achieved good predictive performance on a training dataset and provided the most accurate prediction for wake, which was well validated by an independent sample. HR contributed most to the predictive power of the algorithm. The proposed bedside implementation architecture of the SWB system is being tested and deployed in the NICU of UMC Utrecht. By tracking sleep–wake transitions in clinical practice, the SWB system allows for individualized caregiving activities according to the infant's state in the NICU, thereby improving their sleep quality and protecting their vulnerable developing brain.

Supplementary Material

Supplementary material is available at *SLEEP* online.

Funding

This work was supported by the European Commission, Horizon 2020 Marie Skłodowska-Curie Actions [Grant agreement number: EU H2020 MSCA-ITN-2018-#813483, INtegrating Functional Assessment measures for Neonatal Safeguard (INFANS)].

Disclosure Statement

Financial disclosure: none.

Non-financial disclosure: none.

Acknowledgements

We thank Annemarie van 't Veen, Marieke Bluemink, Giulia de Luca, René van de Vosse, Sander Tan, Chanel Sam, Anne Bik, Marit Knoop and Charlotte Teunis for their contributions to the Sleep Well Baby project. The Sleep Well Baby project originated during Dutch Hacking Health 2019. The authors acknowledge support from the Applied Data Analytics in Medicine (ADAM) program of the UMC Utrecht in the early stages of this project, Tim Pijl and Lars van den Berg from Finaps, data manager Karlijn Eerenberg-Peffer from UMC Utrecht.

Data Availability

Individual participant data that underlie the results reported in this article, after de-identification, can be made available as part of further research collaborations. Interested parties should contact the corresponding author (JD). Any data sharing will be subject to meeting the Privacy Regulations of UMC Utrecht, the General Data Protection Regulation (GDPR) and the General Data Protection Regulation Implementation Act.

References

1. Chawanpaiboon S, *et al*. Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis. *Lancet Glob Health* 2019;**7**(1):e37–e46. doi:10.1016/S2214-109X(18)30451-0.
2. Volpe JJ. Dysmaturation of premature brain: importance, cellular mechanisms, and potential interventions. *Pediatr Neurol*. 2019;**95**:42–66. doi:10.1016/j.pediatrneurol.2019.02.016.
3. Boardman JP, *et al*. Invited Review: Factors associated with atypical brain development in preterm infants: insights from magnetic resonance imaging. *Neuropathol Appl Neurobiol*. 2020;**46**(5):413–421. doi:10.1111/nan.12589.
4. Knoop MS, *et al*. Current ideas about the roles of rapid eye movement and non-rapid eye movement sleep in brain development. *Acta Paediatr*. 2021;**110**(1):36–44. doi:10.1111/apa.15485.
5. Cao M, *et al*. Early development of functional network segregation revealed by connectomic analysis of the preterm human brain. *Cereb Cortex*. 2017;**27**(3):1949–1963. doi:10.1093/cercor/bhw038.
6. Ednick M, *et al*. A review of the effects of sleep during the first year of life on cognitive, psychomotor, and temperament development. *Sleep* 2009;**32**(11):1449–1458. doi:10.1093/sleep/32.11.1449.
7. Georgoulas A, *et al*. Sleep-wake regulation in preterm and term infants. *Sleep* 2021;**44**(1). doi:10.1093/sleep/zsaa148.
8. Prechtl HF. The behavioural states of the newborn infant (a review). *Brain Res*. 1974;**76**(2):185–212. doi:10.1016/0006-8993(74)90454-5.
9. Peirano P, *et al*. Sleep-wake states and their regulatory mechanisms throughout early human development. *J Pediatr*. 2003;**143**(4 Suppl):S70–S79. doi:10.1067/s0022-3476(03)00404-9.
10. del Rio-Bermudez C, *et al*. Active sleep promotes functional connectivity in developing sensorimotor networks. *Bioessays* 2018;**40**(4):e1700234. doi:10.1002/bies.201700234.
11. van den Hoogen A, *et al*. How to improve sleep in a neonatal intensive care unit: A systematic review. *Early Hum Dev*. 2017;**113**:78–86. doi:10.1016/j.earlhumdev.2017.07.002.
12. Uchitel J, *et al*. Early development of sleep and brain functional connectivity in term-born and preterm infants. *Pediatr Res*. 2022;**91**(4):771–786. doi:10.1038/s41390-021-01497-4.
13. Als H, *et al*. The newborn individualized developmental care and assessment program (NIDCAP) with Kangaroo Mother Care (KMC): comprehensive care for preterm infants. *Curr Womens Health Rev* 2011;**7**(3):288–301. doi:10.2174/157340411796355216.
14. Altimier L, *et al*. The neonatal integrative developmental care model: advanced clinical applications of the seven core measures for neuroprotective family-centered developmental care. *Newborn Infant Nurs Rev*. 2016;**16**(4):230–244. doi:10.1053/j.nainr.2016.09.030.
15. Grigg-Damberger MM. The visual scoring of sleep in infants 0 to 2 months of age. *J Clin Sleep Med*. 2016;**12**(3):429–445. doi:10.5664/jcsm.5600.
16. Crowell DH, *et al*. Infant polysomnography: reliability. *Sleep* 1997;**20**(7):553–560.
17. Werth J, *et al*. Unobtrusive sleep state measurements in preterm infants: a review. *Sleep Med Rev*. 2017;**32**:109–122. doi:10.1016/j.smrv.2016.03.005.
18. Koolen N, *et al*. Automated classification of neonatal sleep states using EEG. *Clin Neurophysiol* 2017;**128**(6):1100–1108. doi:10.1016/j.clinph.2017.02.025.
19. de Wel O, *et al*. van Huffel S. Decomposition of a multiscale entropy tensor for sleep stage identification in preterm infants. *Entropy*. 2019;**21**(10):936.
20. Dereymaeker A, *et al*. An automated quiet sleep detection approach in preterm infants as a gateway to assess brain maturation. *Int J Neural Syst*. 2017;**27**(06):1750023.
21. Ansari AH, *et al*. Quiet sleep detection in preterm infants using deep convolutional neural networks. *J Neural Eng*. 2018;**15**(6):66006. doi:10.1088/1741-2552/aadc1f.
22. Ansari AH, *et al*. A convolutional neural network outperforming state-of-the-art sleep staging

algorithms for both preterm and term infants. *J Neural Eng*. 2020;**17**(1):016028.
23. Ansari AH, *et al*. A deep shared multi-scale inception network enables accurate neonatal quiet sleep detection with limited EEG channels. *IEEE J Biomed Health Inf*. 2022;**26**(3):1023–1033. doi:10.1109/JBHI.2021.3101117.
24. Werth J, *et al*. Deep learning approach for ECG-based automatic sleep state classification in preterm infants. *Biomed Signal Proc Control*. 2020;**56**:101663.
25. Mirmiran M. The function of fetal/neonatal rapid eye movement sleep. *Behav Brain Res*. 1995;**69**(1-2):13–22. doi:10.1016/0166-4328(95)00019-p.
26. Marks GA, *et al*. A functional role for REM sleep in brain maturation. *Behav Brain Res*. 1995;**69**(1-2):1–11. doi:10.1016/0166-4328(95)00018-o.
27. de Groot ER, *et al*. The value of cardiorespiratory parameters for sleep state classification in preterm infants: A systematic review. *Sleep Med Rev*. 2021;**58**:101462.
28. Futoma J, *et al*. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health* 2020;**2**(9):e489–e492.
29. de Groot ER, *et al*. Creating an optimal observational sleep stage classification system for very and extremely preterm infants. *Sleep Med*. 2022;**90**:167–175. doi:10.1016/j.sleep.2022.01.020.
30. Biau G, *et al*. A random forest guided tour. *Test*. 2016;**25**(2):197–227.
31. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*. 2006;**27**(8):861–874. doi:10.1016/j.patrec.2005.10.010.
32. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg Uv, Bengio S, *et al*., eds. *Advances in Neural Information Processing Systems*. Vol **30**. Curran Associates Inc.; 2017. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.
33. Reulecke S, *et al*. Autonomic regulation during quiet and active sleep states in very preterm neonates. *Front Physiol*. 2012;**3**:61. doi:10.3389/fphys.2012.00061.
34. Menon AK, Jiang XJ, Vembu S, Elkan C, Ohno-Machado L. Predicting accurate probabilities with a ranking loss. In: Proceedings of the 29th International Conference on International Conference on Machine Learning. International Conference on Machine Learning. Vol **2012**. NIH Public Access; 2012:703. https://dl.acm.org/doi/proceedings/10.5555/3042573.