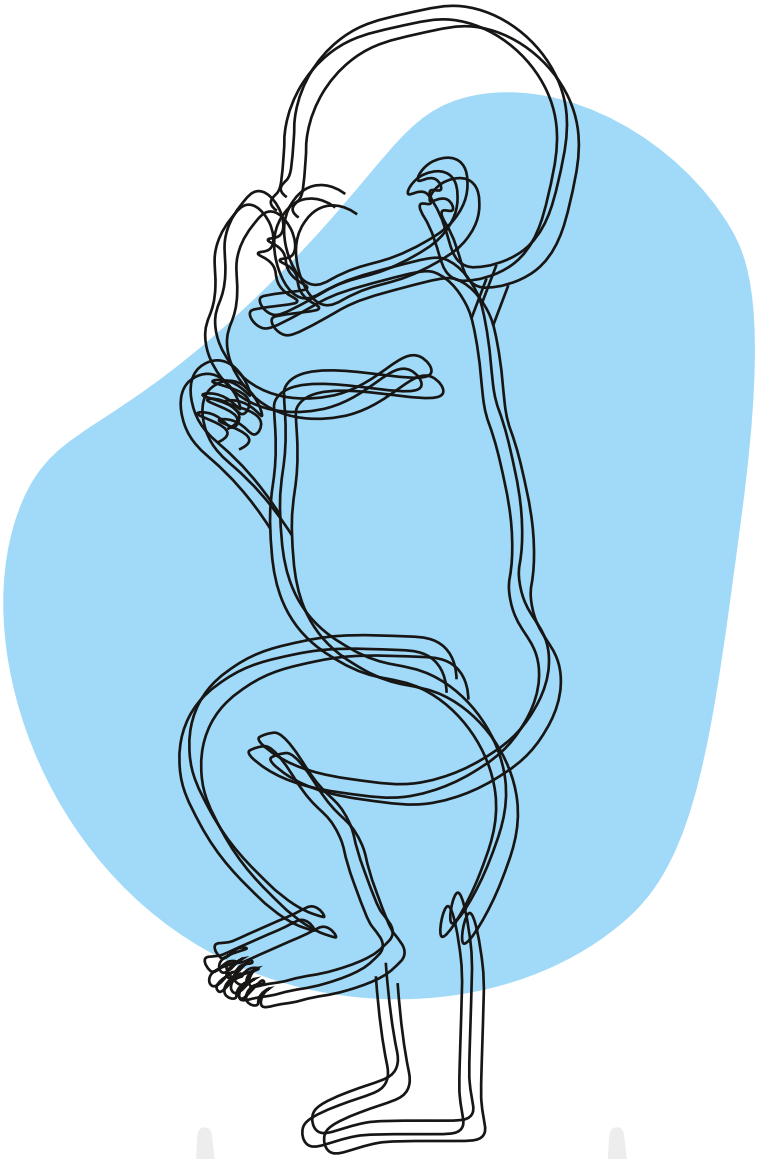


AUTOMATED SLEEP MONITORING BASED ON VITAL SIGNS IN CRITICALLY ILL CHILDREN

MASTER THESIS

ANNE MEESTER

Automated Sleep Monitoring Based on Vital Signs in Critically Ill Children

Anne M. Meester
Student number: 4299132
January 2023

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in
Technical Medicine
Leiden University | Delft University of Technology | Erasmus University Rotterdam

Master thesis project (TM30004; 35 ECTS)
Dept. of Biomechanical Engineering, TU DELFT
Supervisor(s):
dr. ir. D.M.J. Tax
dr. R.C.J. de Jonge
dr. J.W. Kuiper
drs. A.B.G. Cramer
drs. E. van Twist

Thesis committee members:
prof. dr. K.F.M. Joosten (chair)
dr. ir. D.M.J. Tax
dr. R.C.J. de Jonge
dr. J.W. Kuiper

Universiteit Leiden
TU Delft Delft University of Technology
Erasmus Universiteit Rotterdam

Preface

This report represents the final part of my studies at the Delft University of Technology, Erasmus University Rotterdam and Leiden University. In Delft, I initially started studying Applied Mathematics. Unfortunately, to my taste this study lacked practical application of the obtained knowledge, which is why I started in 2015 in the second cohort of the Bachelor in Clinical Technology. After completing this Bachelor within the prescribed three years in 2018, I started with the corresponding Master in Technical Medicine.

During the internships in my Master's program, I had the opportunity to work in different departments within the hospital and contribute to different projects. I was able to develop myself on both a clinical and technical level. I discovered my passion for programming, where I lost track of time more than once, and I discovered that my main interest lies in the intensive care. On the one hand because this was very diverse in terms of clinical experiences and on the other hand because I think there are many possibilities in the technical field. I was therefore delighted that, after an internship in the Neonatal Intensive Care Unit and an internship in the Adult Intensive Care Unit, I was provided the opportunity to start my graduation internship at the Paediatric Intensive Care Unit. Here, I could learn even more about the wide variety of diseases and treatments that occur, combined with doing a project where I could program for whole days on end: the perfect combination.
This project marked my first encounter with machine learning. As this was unfamiliar territory for me, I went back once again to follow additional lectures and to take one last exam. It was a very exciting challenge to start working on and I enjoyed the time on the department.

This Master thesis is the end product of my graduation internship and finalises my time as a student. I am very excited about the future, where I will further develop my knowledge in medicine and technology and apply the combination in practice.

A.M. Meester
Rotterdam, December 2022

Contents

List of abbreviations

AASM	American Association of Sleep Medicine
AUROC	Area under the receiver operator characteristics
CI	Confidence interval
CV	Cross-validation
ECG	Electrocardiogram
EDF	European Data Format
EEG	Electroencephalography
EMG	Electromyography
EOG	Electrooculography
FPR	False positive rate
HF	High-frequency
HR	Heart rate
HRV	Heart rate variability
ICU	Intensive care unit
LF	Low-frequency
LSTM	Long short term memory
MREC	Medical Research Ethics Committee
nCV	Nested cross-validation
NICU	Neonatal intensive care unit
NREM	Non-rapid eye movement
PCA	Principal component analysis
PICU	Paediatric Intensive Care Unit
PSD	Power spectral density
PSG	Polysomnography
PTT	Pulse transit time
REM	Rapid eye movement
ROC	Receiver operator characteristics
SD	Standard deviation
SWT	Stationary wavelet transform
TPR	True positive rate
VLF	Very-low-frequency
XGBoost	Extreme gradient boosting
XML	Extensible Markup Format

Abstract

Introduction: Critically ill children admitted to the Paediatric Intensive Care Unit (PICU) have a high risk of disruption of their normal sleep rhythm, which is associated with disturbances in physiology and negative effects on psychological and cognitive functioning. There is a need for real-time, automatic sleep monitoring to minimise disruptions in sleep patterns. The main objective of this thesis was to develop a machine learning model that can classify sleep based on vital signs in critically ill children. In addition, methods were investigated to optimise the decision criteria in multi-class problems.
Methods: Three machine learning algorithms, logistic regression, random forest and extreme gradient boosting (XGBoost), were developed based on both the combination of features extracted from electrocardiogram (ECG) signal and pulse transit time (PTT) and on ECG features alone. To gain insight into the number of sleep stages that could be distinguished, the models were developed for two-class, three-class, four-class and five-class staging. The models were developed, trained and evaluated on a diagnostic dataset (n = 90) containing polysomnography (PSG) measurements of non-critically ill children. During model development, the decision criteria for the different classes were jointly optimised. To evaluate whether the models were generalisable to the PICU population, external validation was performed on a set of 8 of PICU patients.
Results: For each number of sleep stages, the three models performed similarly. However, there was an increase in performance with a decrease in the number of sleep stages with balanced accuracies varying between 0.70 and 0.72 in two-class staging and between 0.41 and 0.42 in five-class staging. External validation on the PICU dataset showed a markedly worse performance for all three models with balanced accuracies varying between 0.56 and 0.62 in two-class staging and between 0.22 and 0.23 in five-class staging.
Conclusion: Machine learning models for sleep classification in children based on vital signs have been developed and show promising results. Nonetheless, the developed models are not generalisable to the PICU population. Further research is recommended with a focus on improving the models such that they can be applied in the PICU. Combining information extracted from vital signs with EEG signal and developing the models directly on PICU data should be considered to improve the performance and thereby contribute to personalised care and minimising sleep disturbances.

Acknowledgements

I would like to express my sincere gratitude to my supervisors for their continuous guidance and support during this graduation project. Rogier and Jan Willem, thank you for the opportunity to work in your department as a graduate intern, for your involvement during this project and for your extensive feedback. Your positive attitude helped a lot to recognise the added value even with small victories and to be proud of the results. In addition, you made sure that I did not lose sight of my end goal and did not get stuck on certain topics for too long. David, thank you for your help and advice in developing my algorithm. Although I had no knowledge of machine learning at the start of my graduation project, I acquired it in a short time with your help. I always really looked forward to the Mondays when we discussed my results and progress together. These weekly appointments ensured that I thoroughly thought everything through and documented it so well that by the end of my project, I no longer encountered unforeseen problems. Your enthusiasm always provided motivation to explore things further and learn more. Arnout and Eris, thank you for being my first point of contact. As soon as I got stuck somewhere, I was often able to get to you on the same day. Arnout, from the first meeting you managed to enthuse me about sleep research. Your vast knowledge of sleep has ensured that I kept the best interests of the research in mind. You asked challenging questions, which kept me sharp. Eris, I enjoyed having you on the team. When there was a difference in opinion between the technical and the medical side, you were a huge help in balancing the interests. You managed to always look at the bigger picture and gave me good advice. You are definitely an asset to this department! Koen, thank you for participating in my committee. To all colleagues in the department, thank you for taking the time to explain things to me and provide feedback on my clinical performance.

1

Introduction

Sleep plays a critical role in children's growth and development, emotional health and immune function.[1,2] Critically ill children admitted to the Paediatric Intensive Care Unit (PICU) have a high risk of disruption of their normal sleep rhythm due to a chaotic environment, medication, pain associated with the underlying disease, and interruptions by caregivers.[1,3] Frequent awakenings and short sleep durations are associated with disturbances in the immune system and with increased pain perception.[1,3,4] They also affect psychological and cognitive functioning.[1,4] Ideally, real-time sleep monitoring should be used to consider the child's sleep pattern in the scheduling of care and interventions. However, real-time sleep monitoring in children is not yet possible.

The gold standard to classify sleep is polysomnography (PSG).[5] A standard PSG measurement includes at least electroencephalography (EEG), electrooculography (EOG), electrocardiography (ECG), and electromyography (EMG).[6,7] These signals are visually scored by experts that assign a sleep stage to every 30-seconds epoch, according to the American Association of Sleep Medicine (AASM) criteria.[8] These AASM criteria are based on normal EEG signals and are therefore less suitable for use in critically ill patients who often show abnormal EEG patterns.[9,10] Furthermore, PSG entails an additional burden for patients, the manual scoring of these signals is a time-consuming process,[7] and the PSG recordings can only be assessed after the measurements.

Distinction is made between different types of sleep: rapid eye movement (REM) sleep and non-rapid eye movement (NREM) sleep. NREM sleep is subdivided into NREM stage 1 (NREM 1), NREM stage 2 (NREM 2) and NREM stage 3 (NREM 3),[11] where NREM 1 represents the lightest sleep and NREM 3 the deepest sleep.[2] Each type of sleep is associated with unique functions and both neurological and physiological characteristics.[2,11] During NREM sleep there is relatively little brain activity and this stage is characterised by low, regular heartrate and breathing.[2,11] The NREM sleep phase is assumed to primarily have a restorative and rest-enabling function.[4,11] REM sleep is characterised by rapid eye movements, high brain activity and irregular heartrate and breathing.[2,11] It is considered essential for brain development and memory consolidation.[2,4,11] During regular sleep, the stages alternate cyclically. Sleep behaviour changes during the development of a child. New-borns sleep most of the day at irregular times.[2-4,11] With advancing age, both duration and frequency of sleep decrease.[2,3] Given the specific functions of each sleep stage, it is important that all are completed without interruption. Use of real-time automatic sleep monitoring would be a valuable contribution to personalised care, where each patient's sleep pattern can be considered in the scheduling of care and sleep interruptions can be reduced.

In recent years, interest has been aroused in developing automated sleep monitoring methods for adults and preterm infants.[9,12-20] Most of the applied algorithms use features derived from the EEG signal for sleep classification.[12-14,17-19] However, EEGs are not regularly performed on the PICU and adequate use of these algorithms therefore requires additional measurements. Continuous monitoring of various vital signs is already common practice in most of the children admitted to the PICU and the resulting signals therefore provide an opportunity. Several studies have shown that there is a correlation between vital signs and sleep patterns.[21-24] Especially heart rate (HR) and heart rate variability (HRV) show a consistent relationship with sleep.[21-24] In addition, a correlation is observed in literature between sleep and pulse transit time (PTT), [25,26] which is the time duration for an arterial pulse pressure wave to travel from the left ventricle of the heart to a predetermined peripheral site. The PTT is thought to be inversely related to blood pressure.[27,28]

To our best knowledge, no automated sleep classification models based on vital signs have been applied in critically ill children admitted to the PICU. Recent studies have attempted to develop such models for the adult Intensive Care Unit (ICU) and the Neonatal Intensive Care Unit (NICU).[9,15,20] These models, based on machine learning, show promising results. However, the development of machine learning

models on medical data presents challenges. An example of a common issue is that the samples are not evenly distributed over the different classes, resulting in an imbalanced dataset. Machine learning models are trained to distinguish between classes, but it is possible that they are biased towards certain (majority) classes and are therefore not able to adequately distinguish among all classes.

Machine learning models assign a probability score associated to a class to each sample of the dataset.[29] If a model is biased towards a certain class, the corresponding probability score will be higher. By applying a decision criterion, this probability score can be converted into a prediction. Using the default decision criterion, the class with the highest probability score is predicted.[30] However, this does not always lead to an optimal representation of the predicted probabilities. In two-class problems, optimising the decision criterion by changing the operating point is a known solution.[29] In multi-class problems, this is much more complicated, where no proven solution is known to date.[30,31]

The main objective of this thesis was to develop a machine learning model that can classify sleep stages based on vital signs in critically ill children admitted to the PICU. To achieve this, machine learning models have been developed on PSG data for two-class (sleep – wake), three-class (NREM – REM – wake), four-class (NREM 3 – NREM 1&2 – REM – wake) and five-class (NREM 3 – NREM 2 – NREM 1 – REM – wake) staging. It was examined how many sleep stages can be distinguished by the model. In addition to developing and testing different models, this study investigated optimisation of the decision criterion in multi-class problems.

2

Methods

2.1 Study population

For this master thesis, two independent datasets were used. The first dataset was used to develop, train and evaluate the models and was obtained retrospectively from non-critically ill children who underwent a PSG measurement between 2017 and 2022 for diagnostic purposes (e.g., for obstructive sleep apnoea), at the Erasmus MC Sophia Children's Hospital, Rotterdam, the Netherlands. PSG measurements of patients older than 6 months and younger than 18 years were included, where the age of patients born prematurely (<37 weeks gestational age) was corrected until the postnatal age of 2 years. Patients were excluded if more than 25% of one of the required signals was missing from the recording, or quality of data was low due to noise and artefacts and/or data was unreadable. Due to the retrospective study design, informed consent was not required, which was confirmed by the Medical Research Ethics Committee (MREC) of the Erasmus MC.

Considering the changes in EEG over maturation, six age categories were defined:[32] 6-12 months, 1-3 years, 3-5 years, 5-9 years, 9-13 years and 13-18 years. A total of 90 PSG recordings, 15 per age category, were consecutively acquired for the diagnostic dataset.

The second dataset was used for external validation of the models and was prospectively obtained from critically ill children admitted to the PICU of Erasmus MC, Sophia Children's Hospital between 2020 and 2022. These patients participated in either a study investigating the circadian rhythm in children admitted to the PICU (Critical Clock) or a trial investigating the effect of continuous versus intermittent nutrition in PICU patients (ContInNuPIC trial).[33] All children (term born – 18 years) admitted to the PICU with an expected length of stay of at least two days were included in these studies. To be eligible to participate in the ContInNuPIC trial, the children had to be dependent on artificial nutrition. The exclusion criteria differed between the two studies. For the ContInNuPIC trial, the exclusion criteria were possibility to 'oral' feeds, a 'do not resuscitate' code, expected death within 24 hours, re-admission to the PICU after previous randomisation for this trial, transfer from another ICU after a stay of three or more days, ketoacidotic/hyperosmolar coma on admission, metabolic diseased requiring specific diets, premature newborns or short bowel syndrome.[33] For the Critical Clock study, the exclusion criteria were premature newborns, syndrome associated with mental retardation, hydrocortisone use in the three days prior to admission, melatonin use in the 24 hours before admission, transfer from another ICU, weight below two kilograms, expected not to receive an arterial line or previous inclusion in the Critical Clock study. For these trials, prior written consent of all parents was obtained and MREC has provided permission for use of the patient data in this study (MEC-2020-0333 and MEC-2020-0137). All available PSG records from patients between 6 months and 18 years of age were included in the PICU dataset. Both datasets contain patient characteristics obtained from medical records.

2.2 Data acquisition

The measurements differed slightly per dataset. In the diagnostic dataset, all patients underwent a PSG measurement (Brain RT, OSG, Rumst, Belgium), including measurements of: 14-channel EEG, EOG, EMG, ECG, nasal airflow, chest and abdominal wall motion, arterial blood oxygen-haemoglobin saturation (SpO2), transcutaneous partial pressure of carbon dioxide (tcpCO2) and a capillary blood gas test. In the PICU dataset, the measurements were limited to a single-sided 7-channel EEG and regular EOG, EMG and ECG. Since the recordings were collected from BrainRT alone and the measurements in the PICU dataset were limited, these recordings did not contain PTT data.

After the measurements, the recordings were visually scored by a trained clinical neurophysiology technician and one of the five sleep stages was assigned per 30-seconds epoch according to the AASM criteria.[8] For calculation of the interrater agreement, PSG recordings of both datasets were initially scored once by a neurophysiology technician, after which a time span of 3 hours per recording of the PICU dataset was scored a second time by another technician.

A difference in staging between the two datasets was the regular usage of the N stage in the PICU dataset. The N stage was assigned to epochs that had the characteristics of NREM sleep, but where no distinction could be made between NREM 1, NREM 2 and NREM 3 due to atypical EEG characteristics.

The visually scored sleep stages and the raw ECG signal were manually exported from BrainRT in European Data Format (EDF). In addition, Extensible Markup Format (XML) files were created by BrainRT for the diagnostic dataset only, in which PTT per time episode is included.

2.3 Pre-processing and feature extraction

Supervised machine learning algorithms were developed based on a set of features derived from the ECG signal and PTT data.

QRS-detection

To calculate the ECG features, R-peaks were first detected using the Kalidas (2017) method.[34] This method is based on the Pan Tompkins algorithm.[35] However, instead of a bandpass filter, stationary wavelet transform (SWT) is used to remove noise and to emphasise the QRS-peaks.[34,36] Besides this method's good performance, it is also very suitable for real-time QRS detection. The Python script for the method was obtained via *Neurokit*.[37]

Time windows

In practice, a sleep stage is assigned for every 30-seconds epoch, taking contextual information into account. This means that classification of surrounding epochs is used to determine the class of the current epoch. To include contextual information in the algorithm and to ensure sufficient data to capture changes in autonomous activity,[38] information from the surrounding 8 epochs (120 seconds before and 120 seconds after the respective epoch) was included to calculate the features of the current epoch. Therefore, the algorithm assigns a sleep stage to every 30 seconds epoch, while most features are calculated for a 270 second window around this epoch. Exceptions are the ECG time-domain features, which are also calculated over a window of 30 seconds.

Features

The QRS complexes detected in the ECG signal per time window served as input for calculating a set of 99 cardiac features that are previously described in literature.[39-48] To calculate the 12 features derived from the PTT, the PTT values calculated in BrainRT were used as input. Cardiac features were calculated in time, frequency and nonlinear domain, while PTT features were calculated in time domain only. In addition, the age categories have been added as dummy variables. An overview of all features is listed in Table 1. Features were calculated using the Python package *pyhrv*.[49]

To obtain frequency-domain features of the cardiac signal, an estimate of the power spectral density (PSD) of this signal is required. Three different methods known from literature were used: Welch's method, Lomb-Scargle periodogram and autoregressive method.[50-52] Analysis of PSD provides information about how power distributes as a function of frequency. This is used to evaluate the modulation of the autonomic nervous system on the heart.[52,53] The power of three different frequency bandwidths can be distinguished[39]: (1) power in very-low-frequency range (VLF: 0.003-0.04 Hz), (2) power in low-frequency range (LF: 0.04 Hz-0.15 Hz), and (3) power in high-frequency range (HF: 0.15-0.4 Hz). All three approaches for the PSD estimation were used to calculate all frequency-domain features according to these specified frequency bands.

Artefact detection

A simple form of artefact detection has been applied, in which R-R intervals were considered. If an epoch had R-R intervals less than 250 ms (HR > 240 bpm) or larger than 2000 ms (HR < 30 bpm), it was identified as an artefact and removed from the dataset. In addition, empty values were also removed from the dataset.

Table 1. Overview of the calculated features per epoch.

Features	Description
Cardiac features[49] (n = 99)	
Linear time-domain analysis (n = 44)	
NNI-parameters [ms]	Basic statistical measures from a series of normal R-R intervals (NN intervals) including minimum, maximum, range, mean, median, percentiles (5, 10, 25, 75, 90, 95) and interquartile range.[40-43]
Δ NNI-parameters [ms]	Basic statistical measures from a series of NN interval differences including minimum, maximum and mean.[40,41]
*HR-parameters [bpm]**	Basic statistical measures from a series of heart rate (HR) data including minimum, maximum, range, mean, median, percentiles (5, 10, 25, 75, 90, 95) and interquartile range. [40-42,44]
SDNN [ms]	Standard deviation of all NN intervals.[39-41,43-47]
RMSSD [ms]	The root mean square of the successive differences between N-N intervals.[39-41,43,45-47]
SDSD [ms]	Standard deviation of the successive differences between N-N intervals.[40]
pNN [%]	Total number of pairs of adjacent N-N intervals that differ more than x ms, divided by the total number of N-N intervals. Calculated for x = 20 and x = 50.[39-41,45-47]
Linear frequency-domain analysis** (n = 48)	
Peak frequencies [Hz]	Peak frequencies of all frequency bands including VLF, LF, HF.[48]
Absolute power [ms^2]	Absolute power of all frequency bands, including VLF, LF, HF.[39,45-48]
Relative power [%]	Relative power of all frequency bands, including VLF, LF, HF.[48]
Log power [log]	Logarithmic power of all frequency bands, including VLF, LF, HF.[41,43]
Norm power [-]	Normalised power of the LF and HF frequency band.[39,45-47]
LF/HF [-]	Ratio of the low-to-high frequency power.[39,41,45-47]
Total power [ms^2]	Total power over all frequency bands.[43]
Non-linear indices (n = 7)	
SD 1 [ms]	Poincaré plot standard deviation perpendicular to the line of identity.[48]
SD 2 [ms]	Poincaré plot standard deviation along the line of identity.[48]
SD1/SD2 [-]	Ratio of SD1 to SD2.[48]
SampEn[-]	Sample entropy measures the regularity and complexity of a time series.[43,48]
DFA short	Detrended fluctuation analysis, describes short-term fluctuations.[41,48]
DFA long	Detrended fluctuation analysis, describes long-term fluctuations.[41,48]
PTT features (n = 12)	
Statistical measures [ms]	Minimum, maximum, mean, range, median, percentiles (5, 10, 25, 75, 90, 95) and interquartile range
Other (n = 6)	
Age	Age category as dummy variables.

** Calculated over windows for both 30 seconds and 270 seconds*

*** Calculated an estimate of the power spectral density (PSD) using three different methods: Welch's method, the Lomb-Scargle periodogram and the autoregressive method. For all three estimates, all frequency domain features are computed according to the specified frequency bands: VLF [0.00 Hz – 0.04 Hz], LF [0.04 Hz – 0.15 Hz], HF [0.15-0.40 Hz].[54]*

2.4 Dimensionality reduction

Some features are irrelevant or redundant for the classification of sleep stages. In addition, features can be strongly correlated to each other.[55] Using all features as model input often leads to a complex model with performance degradation and overfitting.[56] Therefore, principal component analysis (PCA) was used, where new variables were computed as linear combinations of the original features.[57,58] Using this method, the majority of the variance could be explained with far fewer components. This method reduced the dimensionality of the data while preserving most of the information and variance within the dataset.[55,58] PCA is highly sensitive to variances within the dataset. If features differ in size, the ones with high variance will be dominant. Therefore, features should be standardised before applying PCA.[55,57] Standardisation to Z-scores was performed for each feature by subtracting the mean (μ) from the original value. This difference is divided by the standard deviation (σ).

2.5 Model development, training and evaluation

Three different classifiers were considered for development of the model: logistic regression, random forest and extreme gradient boosting (XGBoost). Logistic regression is a classifier that bases its prediction on the linear combination of features.[59] Random forest and XGBoost are both ensemble methods based on decision trees. A decision tree is a machine learning model that learns by asking if/else questions, dividing data into subgroups and ultimately leading to classification.[60] Random forest constructs several of these decision trees, each trained on a subset of the data, with the final prediction

determined by the majority vote.[56,60,61] XGBoost sequentially constructs decision trees, with each tree learning from the mistakes of its predecessors.[60]

The three classifiers were developed based on the combination of PTT and ECG features and on ECG features alone. The main focus of interest in this study is in monitoring sleep in children admitted to the PICU. While it is preferable to classify sleep stages in as much detail as possible, distinguishing between sleep and wake would also be a valuable contribution. Part of this thesis is to gain insight into how many sleep stages can be distinguished. Therefore, the models were developed for two-class (sleep – wake), three-class (NREM – REM – wake), four-class (NREM 3 – NREM 1&2 – REM – wake) and five-class (NREM 3 – NREM 2 – NREM 1 – REM – wake) staging. This resulted in the development of 24 (3×2×4) different models. Since stage N in the diagnostic dataset was only assigned to a few epochs in one single patient (0.02%), this stage could not be included in model development and therefore these epochs were removed from analysis.

Baseline performance

Prior to optimisation of the models, a baseline performance was determined for each model. This was done by training the models, with their default hyperparameter settings, on the diagnostic dataset using cross-validation (CV). A five-fold grouped stratified split was applied, with grouping on a patient level and stratification on sleep stages. This method ensures that data from one patient can only occur in one fold and that the distribution of samples for each sleep stage is preserved in the folds.[62]

Hyperparameter optimisation, training, and internal validation of the machine learning models

To optimise and train the models, nested cross-validation (nCV) was applied to the entire diagnostic dataset for each model, as illustrated in Figure 1. For both the inner and outer CV loop, a grouped stratified split was applied: a five-fold split in the outer CV loop as described in section 'Baseline performance' and a three-fold split in the inner CV loop.

The inner CV loop was used for hyperparameter optimisation. A parameter grid was set up for each model as shown in Table S2 and Table S3. Using grid search, the models with all possible combinations of hyperparameter settings were trained and tested during the three-fold CV. The outer loop was then used to obtain the training performance, i.e. the expected performance of the models on unseen data.

During the nCV procedure, the aim was to maximise the metric area under the receiver operator characteristic (AUROC). This metric shows the model's ability to differentiate between classes. The performance of the models is analysed for all possible threshold values for all classes, which makes the metric very suitable for use with a skewed class distribution.[60,63,64] Other measures that were obtained for evaluation of the training performance are accuracy, Cohen's kappa, macro-averaged F1 score and the balanced accuracy. To assess potential difference between age categories, additional internal validation was obtained across age categories.

Post-optimisation technique

A skewed class distribution affects the model performance. To compensate for this, the hyperparameter 'class weight' was set to 'balanced' for logistic regression and random forest. This setting adds weights during training that are inversely proportional to the frequencies of the classes in the input data. This class weight hyperparameter is unfortunately not available for XGBoost, which may result in worse performance.

The models have been optimised by maximising the AUROC. Since AUROC evaluates the performance for all possible decision thresholds, a high AUROC value does not directly lead to correct predictions. It could be that the models are able to correctly predict the sleep stages, but that the threshold has been chosen inefficiently. Even though shifting a threshold in binary problems is a well-known solution, it is not directly applicable to multi-class problems.[65]

In multi-class problems, a receiver operating characteristics (ROC) curve is obtained for each class separately by applying 'one versus the rest'. Therefore, if the threshold for one of these ROC curves is shifted, it will have a direct effect on the other predictions. Various methods have been applied to improve the predictions for the XGBoost classifier. An overview of these methods is presented in section 'Supplementary materials 1'. Finally, weights were added to the posterior probabilities of each class by

means of a weight vector, optimising the balanced accuracy. This ensures that every class is considered equally important.[66]

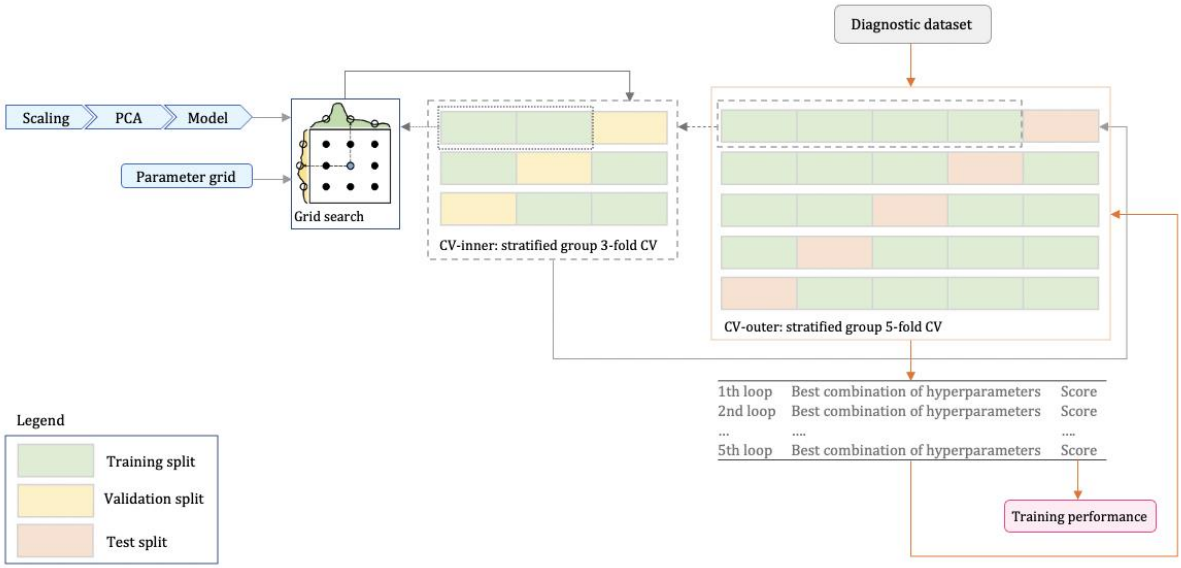


Figure 1. Schematic representation of the nested cross-validation procedure for machine learning model development. The entire diagnostic dataset is divided into five folds in the CV outer loop. Per iteration, the outer training set is used in the CV inner loop, where the data is further divided into three folds. A hyperparameter grid has been set up for each model. The inner training set is fed into a pipeline, where data is successively scaled, PCA is applied and the model is fitted. This pipeline, along with the parameter grid, are fed into the grid search, where all possible parameter combinations are trained on the inner training set and tested on the inner validation set. Validation scores are averaged over the three iterations. The combination of hyperparameters with the highest validation score is returned to the CV outer loop. Here, the model with the hyperparameters is trained on the outer training set and then tested on the outer test set. By performing this for all five iterations and averaging the test scores, a training performance is obtained: the expected performance on unseen data.

External model validation

To investigate whether the models could be generalised to PICU data, external validation was performed. The final optimised models developed based on ECG features alone were fitted once more to the entire diagnostic dataset and then applied to the PICU dataset. Different performance measures were obtained including accuracy, Cohen's kappa, macro-averaged F1 score and balanced accuracy. The 95% confidence intervals (CI) were generated by bootstrapping with replacement across the patients 500 times.

The performance measures were also obtained per PICU patient, to obtain insight in whether the generalisation of the models is valid for the entire PICU population or to specific PICU patients. Since stage N does occur in the PICU data but is not included in the diagnostic data, this might negatively influence the performance. Results were visualised, including how the models classified the unknown N stage. In addition, the results were compared to the interrater agreement, where the same performance measures were obtained as during external validation. To calculate these measures, the labels assigned by one technician were treated as the gold standard and the labels assigned by the other technician were treated as predictions.

Post-optimisation external validation

Since external validation was applied to a dataset acquired from a different patient population, a decrease in performance was to be expected. An experiment was conducted to identify whether any decrease in performance was caused by the model not being able to be generalised well to the PICU population, or whether it was due to the scaling of the model. The post-optimisation procedure was applied once for each model after external validation was completed. A weight vector was obtained for each model, as

described in ‘Supplementary materials 1’, and was multiplied by the posterior probabilities resulting from the model.

2.6 Statistical analysis and software

Pre-processing and development of the models was performed in Python 3.8 using the following packages: NumPy 1.19.2, pandas 1.1.3, Matplotlib 3.3.2, SciPy 1.3.3, pyHRV 0.4.1, NeuroKit2 0.2.0, scikit-learn 1.1.1 and xgboost 0.90.

Baseline characteristics are reported as median (Q1, Q3) or mean (SD) for continuous variables and as percentage (number) for categorical variables. The baseline characteristics were compared between the diagnostic dataset and the PICU dataset using Mann-Whitney U test for continuous variables, Fisher’s exact test for binary categorical variables and chi-squared test for nominal categorical variables. For each analysis, a two-sided p-value less than 0.05 was considered statistically significant.

3

Results

3.1 Patient characteristics

The diagnostic dataset consisted of 90 patients, equally divided over the six predetermined age categories, ranging between 6 months and 17.6 years (median age of 5.0 (2.3-10.5) years). Patient characteristics are presented in Table 2. The PICU dataset consisted of 8 patients with ages ranging between 7 months and 17.3 years (median age of 8.0 (1.0-15.0) years). Gender and age showed no significant differences between the two datasets in contrast to all other considered patient characteristics (i.e. duration of the recording per patient, artefacts, NaN values and sleep stage distribution).

Even though the samples are evenly distributed over the different sleep stages for the diagnostic dataset, 'wake' and 'NREM 1' are considerably less common. As a result, combining stages for development of models with fewer classes leads to a skewed distribution of the data. This is of greater concern when considering the PICU dataset, where a more variable distribution of sleep stages is present.

Table 2. Characteristics of patients included in the diagnostic dataset and included in the PICU dataset.

Variable	**Diagnostic dataset (n = 90)**	**PICU dataset (n = 8)**	**p-value**
Gender, male, % (n)	61.1 % (55)	50 % (5)	0.71*
Median age, years, median (Q1, Q3)	5.0 (2.3, 10.5)	8.0 (1.0, 15.0)	0.34**
PSG / PICU indication, % (n)	Airway obstruction 52.2 % (47) Neuromuscular disease 25.6 % (23) Pulmonary disease 7.8 % (7) Central sleep apnea 2.2 % (2) Unknown 12.2 % (11)	Cardiac failure 12.5% (1) Cardiothoracic surgery 12.5% (1) Sepsis 50.0% (4) Neurological disease 12.5% (1) Oncological disease 12.5% (1)	
Duration of recording per patient, minutes, median (Q1, Q3)	549.0 (494.8, 601.3)	1322.0 (1244.5, 1402.5)	< 0.05**
Total duration of all recordings, hours (n)	857.3 (102873)	171.1 (20592)	
Artefacts in all recordings, % (n)	2.9 % (3003)	1.5 % (313)	< 0.05*
NaN values in all recordings, % (n)	1.2 % (1197)	0.3 % (53)	< 0.05*
Observed sleep stages in all recordings, % (n)			< 0.05***
Wake	14.1% (13910)	33.6 % (6778)	
REM	18.1% (17899)	2.7 % (546)	
NREM 1	10.4 % (10253)	13.0 % (2616)	
NREM 2	27.3 % (26957)	25.4 % (5130)	
NREM 3	30.0 % (29654)	10.8 % (2184)	
N	0.0 % (22)	14.4 % (2909)	

** Fisher's exact test* *** Mann-Whitney U test* **** Chi-squared test*

3.2 Dimensionality reduction

Figure S11 shows that 40 principal components were needed to explain 99% of the variance. For development of the models, the explained variance was set to 99%, resulting in a decrease in dimensionality.

3.3 Baseline performance

The baseline performance of the models using their default hyperparameter settings is presented in Table S4.

3.4 Internal validation of the models

Nested cross-validation

After applying the nCV procedure, the hyperparameters were fixed as shown in Table S2 and Table S3 (Supplementary materials 3). Table 3 provides an overview of the performance metrics of the models. For all models developed based on ECG and PTT features, the AUROC values show that the performance of the model decreases as the number of sleep stages increases. The performance of the three machine learning models is comparable for all assessed numbers of sleep stages. As Table 3 shows, the AUROC values for logistic regression, random forest and XGBoost for two-class staging were 0.82, 0.82 and 0.83 respectively, while for five-class staging they were 0.74, 0.73 and 0.74 respectively.

Similar results were obtained for models developed based on ECG features alone. Once more, a decrease in performance can be observed with an increase in the number of classes. The AUROC values for two-class staging were 0.79 (logistic regression), 0.80 (random forest) and 0.80 (XGBoost), while those for five-class staging were 0.74 (logistic regression), 0.73 (random forest) and 0.74 (XGBoost). The ROC curves presented in section 'Supplementary materials 6' show that in five-class staging, stage NREM 1 and stage NREM 2 are least distinguishable. Section 'Supplementary materials 7' specifically examines the performance of the XGBoost model and further illustrates that NREM 1 and NREM 2 are least distinguishable.

The internal validation results also show that the models developed based on both ECG and PTT features perform slightly better than the models developed based on ECG features only.

A comparison of model performance among all considered age categories showed no significant differences. The AUROC values are presented in Table S5 (Supplementary materials 5).

Table 3. Overview of the performance metrics calculated for internal validation of all models developed based on the combination of ECG and PTT features and ECG features alone. Internal validation was calculated using the nested cross validation procedure on the diagnostic dataset.

	Performance measures					
	ECG and PTT features			**ECG features**		
	LR	RF	XGB	LR	RF	XGB
Two stages						
Accuracy	0.76 (0.03)	0.78 (0.03)	0.77 (0.02)	0.74 (0.04)	0.77 (0.04)	0.73 (0.04)
AUROC	0.82 (0.02)	0.82 (0.02)	0.83 (0.02)	0.79 (0.02)	0.80 (0.03)	0.80 (0.03)
Cohen's kappa	0.32 (0.02)	0.34 (0.02)	0.36 (0.03)	0.28 (0.03)	0.30 (0.03)	0.28 (0.04)
F1 score	0.45 (0.01)	0.46 (0.01)	0.48 (0.03)	0.41 (0.02)	0.43 (0.02)	0.43 (0.04)
Balanced accuracy	0.73 (0.01)	0.73 (0.02)	0.76 (0.02)	0.70 (0.01)	0.71 (0.02)	0.72 (0.02)
Three stages						
Accuracy	0.62 (0.03)	0.61 (0.01)	0.63 (0.01)	0.62 (0.03)	0.61 (0.01)	0.62 (0.03)
AUROC	0.79 (0.03)	0.78 (0.01)	0.80 (0.02)	0.78 (0.03)	0.78 (0.02)	0.79 (0.03)
Cohen's kappa	0.35 (0.04)	0.34 (0.03)	0.36 (0.03)	0.34 (0.04)	0.32 (0.03)	0.35 (0.03)
F1 score	0.55 (0.03)	0.55 (0.02)	0.56 (0.02)	0.54 (0.03)	0.54 (0.02)	0.55 (0.03)
Balanced accuracy	0.61 (0.03)	0.61 (0.02)	0.62 (0.02)	0.59 (0.03)	0.59 (0.03)	0.61 (0.03)
Four stages						
Accuracy	0.48 (0.02)	0.46 (0.02)	0.47 (0.02)	0.47 (0.02)	0.46 (0.01)	0.48 (0.01)
AUROC	0.75 (0.01)	0.74 (0.01)	0.76 (0.01)	0.74 (0.01)	0.74 (0.01)	0.75 (0.01)
Cohen's kappa	0.31 (0.02)	0.28 (0.03)	0.31 (0.03)	0.29 (0.01)	0.28 (0.02)	0.30 (0.02)
F1 score	0.48 (0.02)	0.45 (0.02)	0.46 (0.02)	0.46 (0.02)	0.45 (0.01)	0.47 (0.01)
Balanced accuracy	0.52 (0.01)	0.51 (0.02)	0.52 (0.01)	0.50 (0.00)	0.50 (0.01)	0.51 (0.01)
Five stages						
Accuracy	0.44 (0.03)	0.42 (0.03)	0.45 (0.03)	0.43 (0.03)	0.42 (0.03)	0.45 (0.03)
AUROC	0.74 (0.02)	0.73 (0.02)	0.74 (0.02)	0.74 (0.02)	0.73 (0.02)	0.74 (0.02)
Cohen's kappa	0.29 (0.03)	0.26 (0.04)	0.29 (0.03)	0.27 (0.03)	0.26 (0.03)	0.28 (0.03)
F1 score	0.41 (0.02)	0.39 (0.03)	0.37 (0.03)	0.40 (0.02)	0.39 (0.02)	0.38 (0.02)
Balanced accuracy	0.43 (0.02)	0.41 (0.03)	0.42 (0.01)	0.42 (0.01)	0.41 (0.02)	0.41 (0.01)

Values are presented as mean (standard deviation). LR = logistic regression, RF = random forest, XGB = XGBoost

Post-optimisation technique

Table S6 provides an overview of the balanced accuracies before adding weights to the posterior probabilities and the balanced accuracies after. The balanced accuracies for all XGBoost models improve with the addition of the weights. Section 'Supplementary materials 7' elaborates on this improvement and shows the change of the confusion matrix as a result of adding the new weights. The weights that have been found are presented in Table S7. These will also be applied to the posterior probabilities of the PICU data.

3.5 External validation of the models

The PICU dataset, that was used for external validation of all models, contains data with a different distribution of sleep stages compared to the data from the diagnostic dataset. As presented in Table 2, the amount of wake is higher in the PICU dataset. In addition, less time is spent in REM sleep and stage N only occurs in the PICU dataset.

There is a strong variation in the distribution of sleep per PICU patient, as presented in Figure S17. Patient 1, 5, 6 and 7 experienced all sleep stages present in the diagnostic dataset, only differing in ratios. Patients 2, 3 and 4 spent part of their sleep in the N stage, with patient 2 lacking NREM 1 and NREM 2 sleep and patient 4 only containing N sleep and wake. Patient 8 did not experience REM sleep.

The external validation is both determined for all PICU patients together and for each PICU patient individually. The performances of the external validation across all PICU patients are presented in Table 4 and Figure 2. The three models showed comparable results for all assessed number of sleep stages. The performance of the external validation is worse than of the internal validation for each model, with the largest difference observed in the models with more than three stages (Figure 2).

Table 4. Overview of the performance metrics calculated for external validation of all models developed based on ECG features alone.

	Performance measures			
	LR*	RF*	XGB*	Interrater agreement
Two stages				
Accuracy	0.51 (0.33 – 0.68)	0.64 (0.46 – 0.77)	0.53 (0.31 – 0.72)	0.92
Cohen's kappa	0.13 (0.00 – 0.36)	0.23 (0.00 – 0.49)	0.11 (0.00 – 0.41)	0.79
F1 score	0.51 (0.32 – 0.67)	0.50 (0.30 – 0.68)	0.46 (0.25 – 0.65)	0.90
Balanced accuracy	0.57 (0.45 – 0.69)	0.62 (0.46 – 0.75)	0.56 (0.42 – 0.70)	0.89
Three stages				
Accuracy	0.41 (0.25 – 0.54)	0.54 (0.39 – 0.65)	0.46 (0.30 – 0.61)	0.90
Cohen's kappa	0.10 (0.00 – 0.26)	0.20 (0.00 – 0.38)	0.12 (0.00 – 0.35)	0.78
F1 score	0.33 (0.20 – 0.43)	0.41 (0.29 – 0.50)	0.37 (0.23 – 0.88)	0.82
Balanced accuracy	0.43 (0.31 – 0.54)	0.47 (0.33 – 0.59)	0.46 (0.31 – 0.59)	0.78
Four stages				
Accuracy	0.28 (0.15 – 0.40)	0.30 (0.16 – 0.43)	0.25 (0.14 – 0.37)	0.68
Cohen's kappa	0.06 (0.00 – 0.15)	0.09 (0.00 – 0.22)	0.05 (0.00 – 0.16)	0.58
F1 score	0.17 (0.10 – 0.23)	0.21 (0.13 – 0.29)	0.17 (0.10 – 0.23)	0.68
Balanced accuracy	0.27 (0.17 – 0.34)	0.28 (0.19 – 0.37)	0.27 (0.17 – 0.35)	0.70
Five stages				
Accuracy	0.26 (0.15 – 0.39)	0.26 (0.15 – 0.38)	0.25 (0.14 – 0.37)	0.63
Cohen's kappa	0.05 (0.00 – 0.16)	0.06 (0.00 – 0.18)	0.05 (0.00 – 0.16)	0.54
F1 score	0.14 (0.09 – 0.20)	0.16 (0.10 – 0.23)	0.15 (0.09 – 0.20)	0.59
Balanced accuracy	0.22 (0.14 – 0.29)	0.22 (0.14 – 0.30)	0.23 (0.15 – 0.29)	0.60

LR = logistic regression, RF = Random forest, XGB = XGBoost

**Performance metrics are illustrated with 95% CI values between brackets, calculated using bootstrapping with replacement across the patients 500 times.*

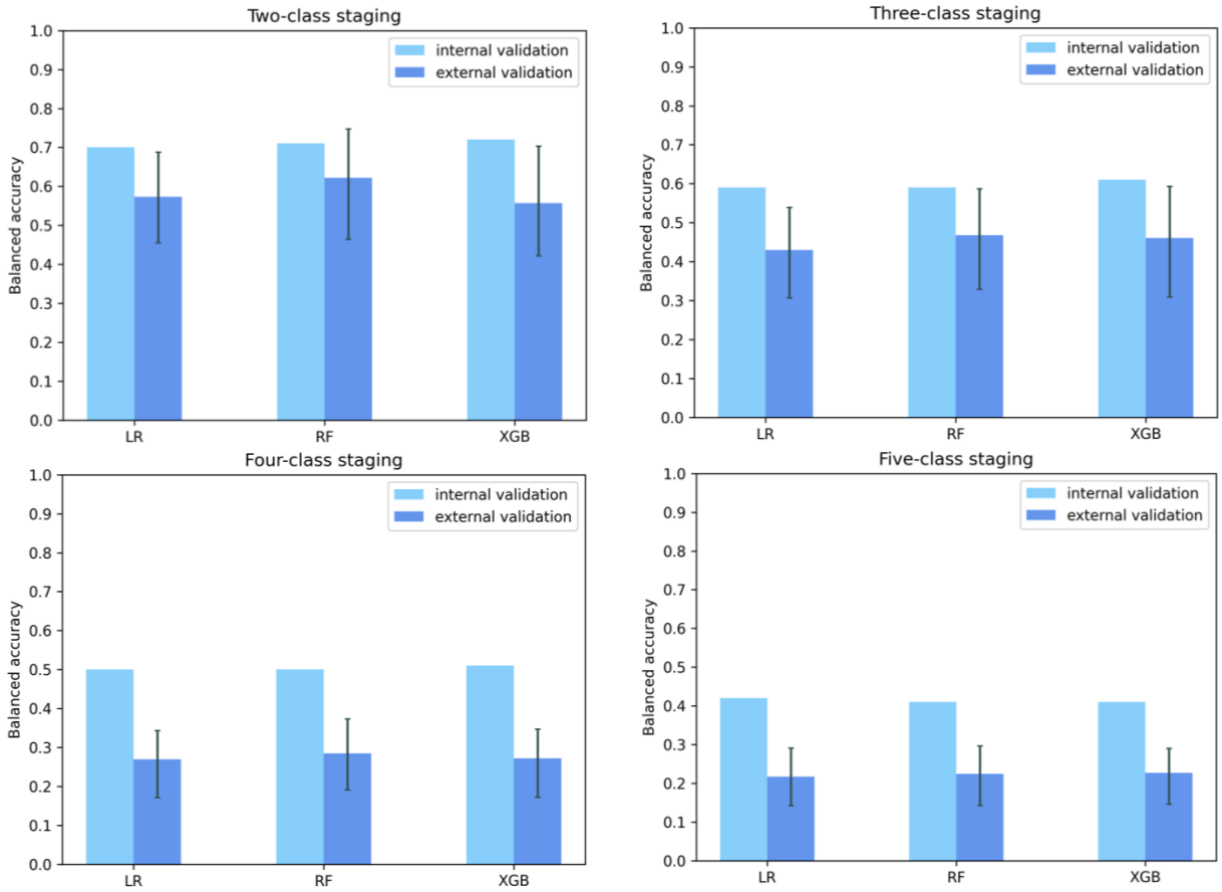


Figure 2. Balanced accuracy of internal and external validation for all models. The confidence intervals for external validation obtained with bootstrapping are illustrated by the error bars.

The assessments of individual PICU patients showed that there were large variations in performance between patients. This is illustrated in Figure 3, which shows the balanced accuracies of the models for four-class staging compared to the interrater agreement.

The following subsections further elaborate on patient 2 and patient 4 of the PICU dataset with the highest and lowest balanced accuracy, respectively, considering the random forest model developed for four-class staging.

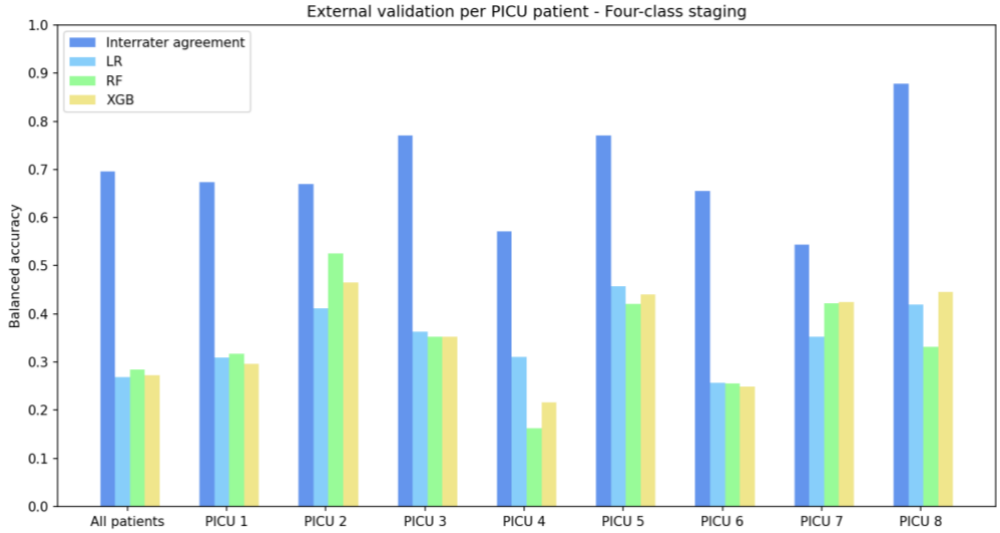


Figure 3. Balanced accuracy per PICU patient for all models for two-class and four-class staging compared to the interrater agreement. (LR = logistic regression, RF = random forest, XGB = XGBoost).

PICU patient 2

The highest balanced accuracy for four-class staging was achieved in PICU patient 2, with a value of 0.52. This is still lower than the interrater agreement, with a calculated balanced accuracy of 0.67. The actual hypnogram and the predicted hypnogram are shown in Figure 4. The confusion matrix of the actual stages with respect to the predicted stages and the confusion matrix of the two technicians with respect to each other are shown in Figure 5. The confusion matrices show similar results for the stages NREM 3, REM and wake, where wake is correctly predicted 98% of the time. However, REM is also regularly predicted as wake. In 48% of the time, the model predicts the N stage as NREM 3. The confusion matrix in Figure 5b shows that in the event of disagreement, if one of the technicians assigns stage N, the other regularly assigns NREM 3.

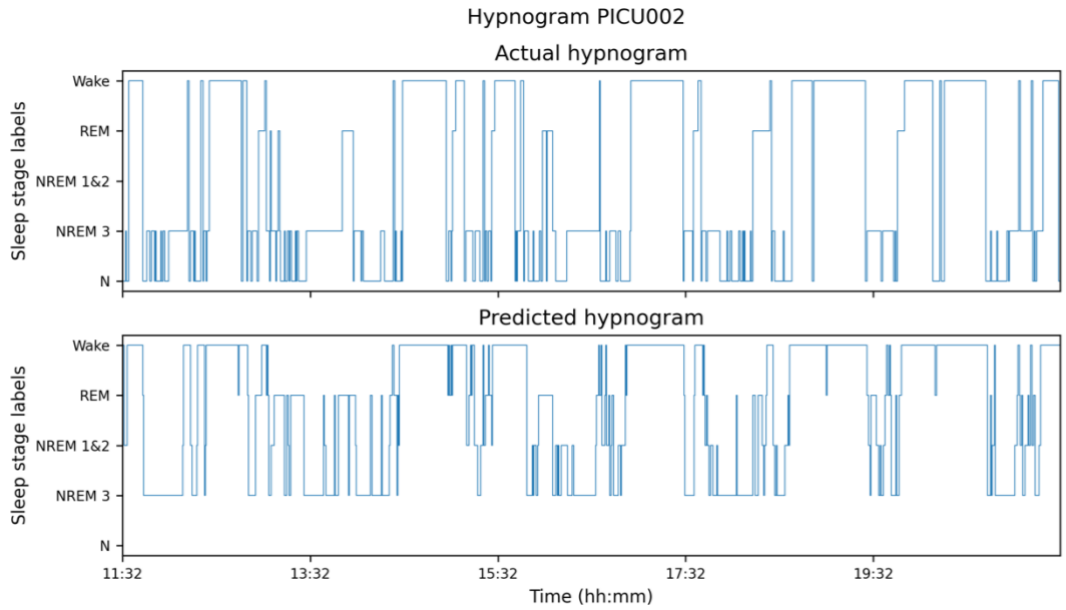
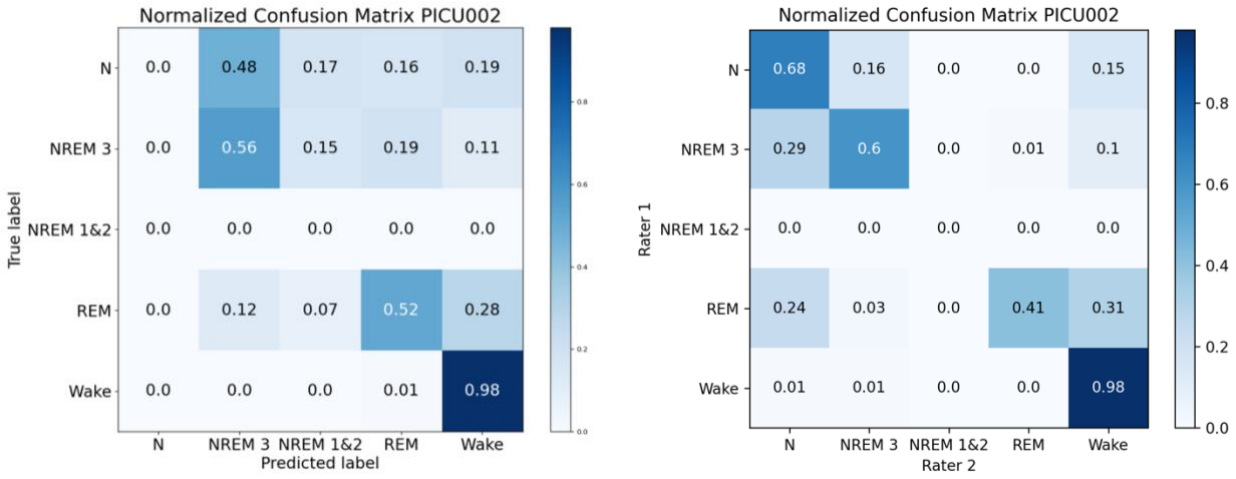


Figure 4. Hypnogram of PICU patient 2, showing the observed and predicted sleep stages over time, determined with the random forest classifier.



(a) Random forest model *(b) Interrater agreement*

Figure 5. Confusion matrices for PICU patient 2

PICU patient 4

The lowest balanced accuracy for four-class staging was achieved in PICU patient 4 with a value of 0.16. This strongly differs from the interrater agreement where a balanced accuracy of 0.57 was found. Patient 4 spent most of the time in stage N and the remainder in wake. The actual hypnogram and the predicted hypnogram are shown in Figure 6. The confusion matrix of the actual stages with respect to the predicted stages and the confusion matrix of the two technicians with respect to each other are shown in Figure 7. The model classifies most of the recording as wake or NREM 3. The epochs assigned to stage N are half of the time classified as wake and the other half as NREM 3. Note that a part of the stages that were labelled as stage N by one of the technicians were also assessed as wake or NREM 3 by the other technician.

The confusion matrices over all patients, as presented in Figure S18, indicate that the technicians are mostly in agreement on the assignment of stage N. In the event of disagreement, often one of the technicians assigns stage N while the other assigns NREM 3 or wake.

The hypnograms for all PICU patients obtained with the random forest model for four-class staging are presented in Supplementary materials 10.

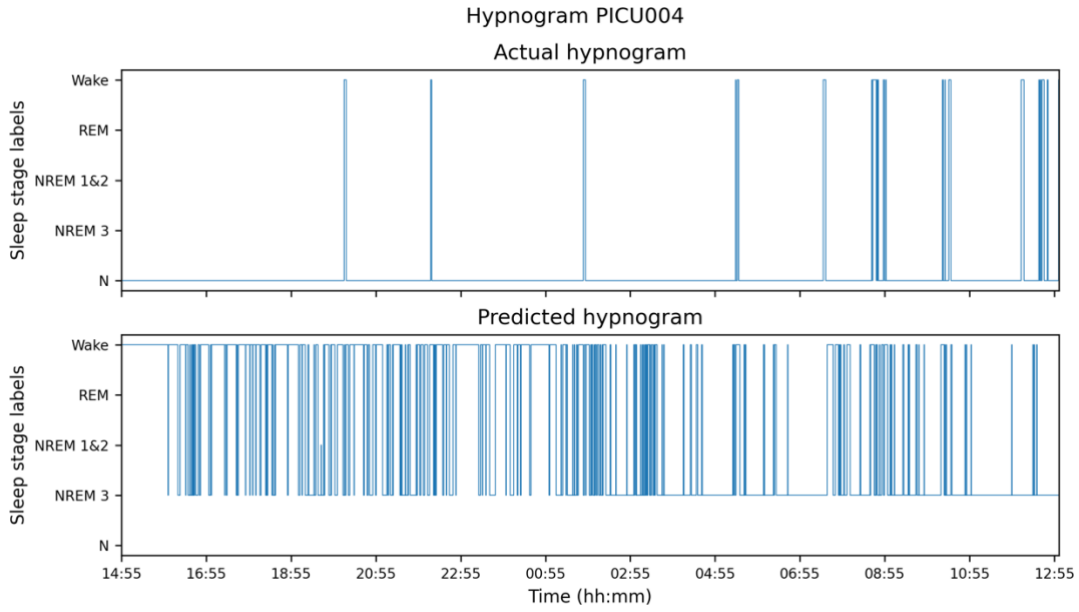
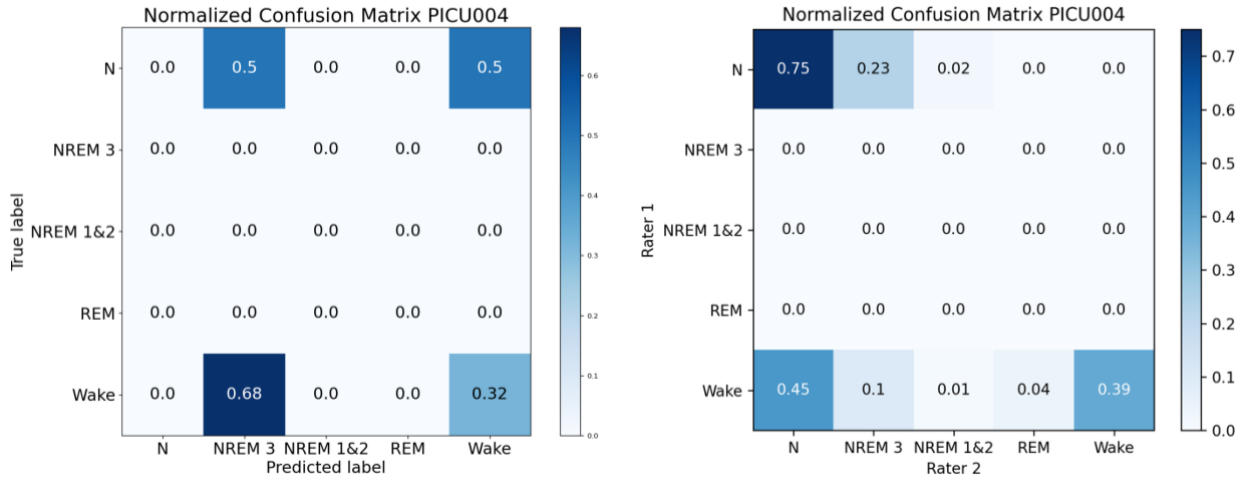


Figure 6. Hypnogram of PICU patient 4, showing the observed and predicted sleep stages over time, determined with the random forest classifier.



(a) Random forest model *(b) Interrater agreement*

Figure 7. Confusion matrices for PICU patient 4

Post-optimisation external validation

Applying the post-optimisation procedure once for each model, after external validation was completed, resulted in the balanced accuracies as presented in Table S8 and Figure S19. The overall balanced accuracies increase slightly in each model, with the largest increase in the logistic regression model and the smallest increase in the random forest model. For two-class staging, the balanced accuracies increased after post-optimisation from 0.57 to 0.61 (logistic regression) and from 0.62 to 0.63 (random forest). For four-class staging, the balanced accuracies increased from 0.27 to 0.32 (logistic regression) and from 0.28 to 0.30 (random forest).

Figure S20 illustrates large variation in balanced accuracies between PICU patients, where optimisation in patients 2, 3 and 6 results in a slightly higher balanced accuracy, whereas this results in a decrease in the other patients.

4

Discussion

The present study aimed to develop machine learning models based on both a combination of ECG signal and PTT and ECG signal alone for sleep classification in children. These models were developed for two-class, three-class, four-class and five-class staging to examine how many sleep stages could be distinguished. The classifiers logistic regression, random forest and XGBoost were considered, each showing comparable performance. The models were able to accurately distinguish between two stages (sleep – wake) and three stages (NREM – REM – wake). However, differentiating between more stages resulted in reduced model performance. External validation on a PICU dataset showed that none of the models were generalisable to the PICU population.

In addition to developing the machine learning models, this study aimed to optimise decision criteria in multi-class problems. By determining and applying a weight vector to the posterior probabilities of the XGBoost model, the poorly calibrated model was improved, resulting in better predictions.

The developed models showed good performance especially for two-class (AUROC range 0.79 – 0.80) and three-class staging (AUROC range 0.78 – 0.79). However, this performance was worse compared to the interrater agreement we found, and the interrater agreement as described in literature.[67,68] To our current knowledge, no automated sleep classification models based on vital signs have been developed for critically ill children. The predictive performance of the models we developed is comparable to models developed based on vital signs in preterm infants.[9,15] However, these are inferior to similar models developed for adults.[41,69] Research by Fonceca et al.[41] showed that four sleep stages could be distinguished with a linear discriminant classifier developed based on cardiac and respiratory features with a Cohen's kappa coefficient of 0.49 and an accuracy of 0.69. This shows better performance than the models developed for this thesis. However, their algorithm was developed based on respiratory features as well, potentially contributing additional information. Furthermore, a subject specific Z-score was used for normalisation, which is not possible if the algorithm needs to be used in real time. The deep neural networks developed by Sun et al.[69] could be used in real-time and show better performance in distinction of the five sleep stages compared to our models. This makes deep neural networks a promising solution for the classification problem. However, they are more difficult to interpret than the machine learning models considered in our study.

Algorithms developed based on EEG features show better results for both preterm infants and children as well as for adults compared to the models developed for this thesis.[13,19,70] Although Zhao et al.[19] still showed the lowest accuracy for prediction of NREM 1 sleep, algorithms developed based on EEG features might be more capable to distinguish the NREM substages from other stages. However, these algorithms entail additional and invasive measurements and thus burden the patient. In contrast, the algorithms developed for this thesis do not require additional measurements, because they are based on vital signs that are already routinely measured in the PICU.

When examining the number of sleep stages in model performance, it was found that fewer stages (two-class staging or three-class staging) could be accurately distinguished and that the performance of the models decreased as the number of sleep stages increased. The models developed for five-class staging indicate that NREM 1 and NREM 2 are poorly distinguishable from the other sleep stages. NREM 1 was often misclassified as REM or wake, while epochs belonging to NREM 2 were often misclassified as one of the other sleep stages, mostly as NREM 3. This can partly be attributed to the method of scoring.[71] According to the guidelines, epochs following an epoch assigned to NREM 3 are scored as NREM 2 if they no longer meet the criteria of NREM 3 and do not contain specific characteristics of wake or REM. Furthermore, after arousal or major body movements, NREM 1 is often assigned to the epoch, unless it meets the criteria of one of the other stages.[71] Scoring of arousals is not included in our models, but they are usually accompanied by an increase in sympathetic activity, which causes changes in HRV,[72,73] potentially contributing to the misclassification of these NREM 1 epochs.

HR and HRV are regulated by the autonomic nervous system. During sleep, autonomic heart control fluctuates between sympathetic and parasympathetic predominance.[74] The sympathetic nervous system predominates during REM sleep and wake. During NREM sleep, when sleep depth increases, activation of the sympathetic nervous system gradually decreases with gradual increase in parasympathetic activity. This leads to change in HR and HRV.[72-74] Different HRV measures therefore show comparable values during wake, REM and NREM 1 sleep.[72] This is a probable explanation of why the models regularly misclassified epochs belonging to NREM 1 as REM or wake.

Due to the frequent misclassification of NREM 1 and NREM 2 with the current models, only those developed for two-class staging and three-class staging can make accurate predictions. Sleep classification based on ECG alone appears unable to properly distinguish NREM 1 and NREM 2.

During internal validation, an adequate AUROC was obtained for the XGBoost model, while the sleep stages were not predicted well. This implied that although the model was able to distinguish the stages, it was not properly calibrated. This could be caused by the skewed distribution of the dataset. To achieve better predictions, two methods were proposed in this report that were applied to both the posterior probabilities obtained from the XGBoost model on the diagnostic dataset and to the posterior probabilities obtained from the logistic regression model applied to different synthetic datasets. The best method was to weigh each class probability, where many combinations of numbers between 10^{-2} and 10 were tried for the weights. The one leading to the highest balanced accuracy was used. This resulted in better predictions and a better-calibrated model. The artificial datasets showed that this had a clear added value, especially in case of a very unbalanced distribution. However, the method becomes computationally more expensive with an increase in the number of classes that need to be distinguished.

When investigating the applicability of the models to the PICU population, the models were applied to the PICU dataset during external validation, resulting in a greatly reduced overall performance compared to internal validation. The post-optimisation procedure as described in section 'Post-optimisation external validation' was applied, whereby the posterior probabilities were multiplied by a weight vector derived for each model, resulting in minimal increase in performance. However, since the weights were determined on the same data on which the performance measures were subsequently calculated, a small improvement in performance could be explained by overfitting. This implies that although the models can classify sleep well, they cannot do so in the PICU population. Several reasons can be put forward to explain why the models based on ECG features alone are currently not applicable in the PICU.

The PICU comprises a heterogeneous population with a great diversity of patients, ages, diagnoses and treatments. The sleep physiology of PICU patients is complex and there are many challenges in monitoring sleep. For example, critical illness and use of sedative medication can impede interpretation of an EEG.[75,76] During critical illness, autonomic dysfunction may occur, which is associated with changes in HRV parameters. In addition, commonly used medications in the ICU, such as chronotropic, antiarrhythmic and antihypertensive drugs, also cause changes in HRV parameters.[77-79] This combination complicates classification of sleep in the PICU population. The effects can be seen in the large variation in performance of our models among PICU patients and the difference in classification between technicians. It is questionable whether the ECG signal, which is strongly affected by the various factors on the PICU, is the best method for sleep classification.

For model development, patients up to six months old have been excluded, because substages of NREM sleep cannot always be distinguished for these patients, resulting in the assignment of epochs to the N stage. The N stage is assigned to epochs that have the characteristics of NREM sleep, but where no distinction can be made between NREM 1, NREM 2 and NREM 3 due to atypical EEG characteristics. During external validation, the N stage was still introduced, regularly occurring in the PICU dataset but not included in the diagnostic dataset. This implies that the interpretation of the EEG in a critically ill child is indeed complicated. Since stage N was not included in training of the models, the models never classified an epoch as stage N. Remarkably, stage N was usually predicted by the models as NREM 3 or wake. This is consistent with how the technicians scored stage N relative to each other, which is usually as NREM 3. This is remarkable because, given the definition of stage N, it would be expected that in the event of disagreement between the technicians, classification would be equally distributed among NREM 1, NREM 2 and NREM 3.

Several limitations of this study should be noted. First, the use of PCA reduces the interpretability of the contributing features. Since new variables are calculated as a linear combination of the original features, no insight can be obtained regarding which features are most distinctive. However, since the algorithm tested on the PICU population now only contains features extracted from the ECG signal, it is less important to know exactly which ECG features contribute the most. In addition, automatic sleep monitoring will not be used to make acute decisions in treatment. Therefore, model performance is assumed of greater importance than model interpretability.

A second limitation is that patients up to six months old were not included in this study, as substages of NREM sleep cannot always be scored before this age. However, this age group concerns about half of the patients admitted to the PICU of the Erasmus MC.

Another limitation is that the model could only be developed based on signals measured by BrainRT. Ideally, the signals would be acquired from the Dräger monitor used routinely for monitoring in the PICU, to enable direct implementation. Since this Dräger monitor was usually not connected during the PSG measurements in the diagnostic dataset, these signals could not be used for model development. As a result, the study was limited to only the signals available in BrainRT. Therefore, information from respiration, for example, which shows promising results in literature, has not been included in the development of the models. Furthermore, PTT was not measured in PICU patients, resulting in the final model only containing features obtained from the ECG signal, while the models developed on the combination of features extracted from ECG and PTT performed slightly better.

Finally, there was not enough PICU data available for development of the machine learning models. Therefore, we opted to use PSG data from non-critically ill children for model development. However, as previously explained, there is a large difference between PICU data and diagnostic data due to, among other things, the large difference in distribution of the sleep stages between the datasets, variation among patients within the PICU dataset and the stage N that occurs regularly in the PICU dataset and was not included in the diagnostic dataset. This makes the diagnostic dataset unsuitable for developing such a model. Nonetheless, it remains valuable research because of the abundance in PSG data and the knowledge we gathered from development of the models. In addition, it is highly recommended to introduce such a model to non-critically ill children because it can assist in manual scoring of the signals, which is a time-consuming and expensive process. This thesis shows that well-performing models have been developed based on the diagnostic dataset, which implies high potential in automatic sleep monitoring based on vital signs.

Automatic real-time sleep monitoring based on vital signs would be a valuable contribution to personalised care, where each patient's sleep pattern can be considered in the care process and sleep disturbances can be reduced. HR and HRV parameters provide a useful basis for automatic sleep classification. ECG signals contain information regarding sleep and are already measured continuously in most patients admitted to the PICU. In the future, it could be interesting to include other routinely measured signals, such as respiration, in the models. Furthermore, adding information extracted from (single-channel) EEG presents an opportunity. Particularly because this would add information other than vital signs and EEG is potentially better able to distinguish between the various NREM stages. Within Erasmus MC, research has been carried out into automatic EEG-based sleep monitoring, showing promising results.[32] Combining these studies marks a first step towards improving classification performance.

Besides the addition of other signals, it is interesting to consider models that make predictions based on time-series data, such as long short term memory (LSTM) networks. With the current data processing, where features are calculated from the signals, time information is extracted from the signal and thus LSTM is not applicable. However, it poses a promising model for the current classification problem.

In the future, sleep classification models for real-time monitoring of PICU patients should be developed on large datasets obtained from PICU patients. Collaboration with other medical centres can contribute to an enlarged dataset and the creation of an external validation set. Furthermore, it would be interesting to investigate in which PICU patients sleep monitoring can be effective. This requires a study including patients with different diagnoses and medications. The classification performance can then be linked to patient characteristics.

Conclusion

In summary, three machine learning models for sleep classification in children based on vital signs have been developed using a diagnostic dataset. In addition, a method has been introduced for optimising decision criteria in multi-class problems for poorly calibrated models. By multiplying each posterior probability by a weight, the balanced accuracy was optimised, leading to better predictions and a better calibrated model.

The developed models achieved good performance on the diagnostic dataset, with NREM 3, REM and wake being distinguished best and increasing model performance as more sleep stages were merged. However, external validation showed that the current models cannot be applied to the PICU population. Further research is recommended with a focus on improving the models such that they can be applied in the PICU population and on combining multiple signals to obtain a better performance.

References

1 Stremler, R. *et al.* Objective Sleep Characteristics and Factors Associated With Sleep Duration and Waking During Pediatric Hospitalization. *JAMA Netw Open* **4**, e213924, doi:10.1001/jamanetworkopen.2021.3924 (2021).

2 Davis, K. F., Parker, K. P. & Montgomery, G. L. Sleep in infants and young children: Part one: normal sleep. *Journal of Pediatric Health Care* **18**, 65-71, doi:https://doi.org/10.1016/S0891-5245(03)00149-4 (2004).

3 Kudchadkar, S. R., Aljohani, O. A. & Punjabi, N. M. Sleep of critically ill children in the pediatric intensive care unit: a systematic review. *Sleep Med Rev* **18**, 103-110, doi:10.1016/j.smrv.2013.02.002 (2014).

4 Carno, M. A. & Connolly, H. V. Sleep and sedation in the pediatric intensive care unit. *Crit Care Nurs Clin North Am* **17**, 239-244, doi:10.1016/j.ccell.2005.04.005 (2005).

5 Crabtree, V. M. & Williams, N. A. Normal Sleep in Children and Adolescents. *Child and Adolescent Psychiatric Clinics of North America* **18**, 799-811, doi:https://doi.org/10.1016/j.chc.2009.04.013 (2009).

6 Jafari, B. & Mohsenin, V. Polysomnography. *Clin Chest Med* **31**, 287-297, doi:10.1016/j.ccm.2010.02.005 (2010).

7 Huang, X., Shirahama, K., Li, F. & Grzegorzek, M. Sleep stage classification for child patients using DeConvolutional Neural Network. *Artif Intell Med* **110**, 101981, doi:10.1016/j.artmed.2020.101981 (2020).

8 Iber, C. The AASM manual for the scoring of sleep and associated events: Rules. *Terminology and Technical Specification* (2007).

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

10 Boyko, Y., Jennum, P. & Toft, P. Sleep quality and circadian rhythm disruption in the intensive care unit: a review. *Nat Sci Sleep* **9**, 277-284, doi:10.2147/nss.S151525 (2017).

11 Bathory, E. & Tomopoulos, S. Sleep Regulation, Physiology and Development, Sleep Duration and Patterns, and Sleep Hygiene in Infants, Toddlers, and Preschool-Age Children. *Current Problems in Pediatric and Adolescent Health Care* **47**, 29-42, doi:https://doi.org/10.1016/j.cppeds.2016.12.001 (2017).

12 Koolen, N. *et al.* Automated classification of neonatal sleep states using EEG. *Clinical Neurophysiology* **128**, 1100-1108, doi:https://doi.org/10.1016/j.clinph.2017.02.025 (2017).

13 Ansari, A. H. *et al.* A convolutional neural network outperforming state-of-the-art sleep staging algorithms for both preterm and term infants. *Journal of Neural Engineering* **17**, 016028, doi:10.1088/1741-2552/ab5469 (2020).

14 Piryatinska, A. *et al.* Automated detection of neonate EEG sleep stages. *Comput Methods Programs Biomed* **95**, 31-46, doi:10.1016/j.cmpb.2009.01.006 (2009).

15 Sentner, T. *et al.* The Sleep Well Baby project: an automated real-time sleep–wake state prediction algorithm in preterm infants. *Sleep*, doi:10.1093/sleep/zsac143 (2022).

16 Cabon, S. *et al.* Audio- and video-based estimation of the sleep stages of newborns in Neonatal Intensive Care Unit. *Biomedical Signal Processing and Control* **52**, 362-370, doi:https://doi.org/10.1016/j.bspc.2019.04.011 (2019).

17 Reinke, L., van der Hoeven, J. H., van Putten, M. J. A. M., Dieperink, W. & Tulleken, J. E. Intensive care unit depth of sleep: proof of concept of a simple electroencephalography index in the non-sedated. *Critical Care* **18**, R66, doi:10.1186/cc13823 (2014).

18 Tsinalis, O., Matthews, P. M., Guo, Y. & Zafeiriou, S. Automatic sleep stage scoring with single-channel EEG using convolutional neural networks. *arXiv preprint arXiv:1610.01683* (2016).

19 Zhao, S. *et al.* Evaluation of a Single-Channel EEG-Based Sleep Staging Algorithm. *Int J Environ Res Public Health* **19**, doi:10.3390/ijerph19052845 (2022).

20 Ganglberger, W. *et al.* Sleep in the Intensive Care Unit through the Lens of Breathing and Heart Rate Variability: A Cross-Sectional Study. *medRxiv*, 2021.2009.2023.21264039, doi:10.1101/2021.09.23.21264039 (2021).

21 Boudreau, P., Yeh, W. H., Dumont, G. A. & Boivin, D. B. Circadian variation of heart rate variability across sleep stages. *Sleep* **36**, 1919-1928, doi:10.5665/sleep.3230 (2013).

22 Fink, A. M., Bronas, U. G. & Calik, M. W. Autonomic regulation during sleep and wakefulness: a review with implications for defining the pathophysiology of neurological disorders. *Clinical autonomic research : official journal of the Clinical Autonomic Research Society* **28**, 509-518, doi:10.1007/s10286-018-0560-9 (2018).

23 Burgess, H. J., Trinder, J., Kim, Y. & Luke, D. Sleep and circadian influences on cardiac autonomic nervous system activity. *Am J Physiol* **273**, H1761-1768, doi:10.1152/ajpheart.1997.273.4.H1761 (1997).

24 de Zambotti, M., Trinder, J., Silvani, A., Colrain, I. M. & Baker, F. C. Dynamic coupling between the central and autonomic nervous systems during sleep: A review. *Neuroscience & Biobehavioral Reviews* **90**, 84-103, doi:https://doi.org/10.1016/j.neubiorev.2018.03.027 (2018).

25 Nisbet, L. C. *et al.* Preschool children with obstructive sleep apnea: the beginnings of elevated blood pressure? *Sleep* **36**, 1219-1226, doi:10.5665/sleep.2890 (2013).

26 Bassam, A. *et al.* Nocturnal dipping of heart rate is impaired in children with Down syndrome and sleep disordered breathing. *Sleep Med* **81**, 466-473, doi:10.1016/j.sleep.2021.03.020 (2021).

27 Vlahandonis, A. *et al.* Pulse transit time as a surrogate measure of changes in systolic arterial pressure in children during sleep. *Journal of Sleep Research* **23**, 406-413, doi:https://doi.org/10.1111/jsr.12140 (2014).

28 Mukkamala, R. *et al.* Toward Ubiquitous Blood Pressure Monitoring via Pulse Transit Time: Theory and Practice. *IEEE Trans Biomed Eng* **62**, 1879-1901, doi:10.1109/tbme.2015.2441951 (2015).

29 Esposito, C., Landrum, G. A., Schneider, N., Stiefl, N. & Riniker, S. GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning. *Journal of Chemical Information and Modeling* **61**, 2623-2640, doi:10.1021/acs.jcim.1c00160 (2021).

30 Lachiche, N. & Flach, P. *Improving Accuracy and Cost of Two-Class and Multi-Class Probabilistic Classifiers Using ROC Curves*. Vol. 1 (2003).

31 Bourke, C., Deng, K., Scott, S. D., Schapire, R. E. & Vinodchandran, N. V. On reoptimizing multi-class classifiers. *Machine Learning* **71**, 219-242, doi:10.1007/s10994-008-5056-8 (2008).

32 Hiemstra, F. W. *Automated EEG-based monitoring in critically ill children*, Delft University of Technology, (2021).

33 Veldscholte K, C. A., de Jonge R, Eveleens R, Joosten K, Verbruggen S. Continuous Versus Intermittent Nutrition in Pediatric Intensive Care Patients: Protocol for a Randomized Controlled Trial. *JMIR Res Protoc* **11(6):e36229**, doi:10.2196/36229 (2022).

34 Kalidas, V. & Tamil, L. *Real-time QRS detector using Stationary Wavelet Transform for Automated ECG Analysis*. (2017).

35 Pan, J. & Tompkins, W. J. A Real-Time QRS Detection Algorithm. *IEEE Transactions on Biomedical Engineering* **BME-32**, 230-236, doi:10.1109/TBME.1985.325532 (1985).

36 Porr, B. & Howell, L. R-peak detector stress test with a new noisy ECG database reveals significant performance differences amongst popular detectors. *bioRxiv*, 722397, doi:10.1101/722397 (2019).

37 Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., & Chen, S. A. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods* **53**, 1689-1696, doi: https://doi.org/10.3758/s13428-020-01516-y (2021).

38 Malik, M. *et al.* Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *European Heart Journal* **17**, 354-381, doi:10.1093/oxfordjournals.eurheartj.a014868 (1996).

39 Francesco, B. *et al.* Linear and nonlinear heart rate variability indexes in clinical practice. *Computational and mathematical methods in medicine* **2012**, 219080-219080, doi:10.1155/2012/219080 (2012).

40 Radha, M. *et al. LSTM knowledge transfer for HRV-based sleep staging*. (2018).

41 Fonseca, P. *et al.* Sleep stage classification with ECG and respiratory effort. *Physiol Meas* **36**, 2027-2040, doi:10.1088/0967-3334/36/10/2027 (2015).
42 Willemen, T. *et al.* An evaluation of cardiorespiratory and movement features with respect to sleep-stage classification. *IEEE J Biomed Health Inform* **18**, 661-669, doi:10.1109/jbhi.2013.2276083 (2014).
43 Lucchini, M., Pini, N., Fifer, W. P., Burtchen, N. & Signorini, M. G. Entropy Information of Cardiorespiratory Dynamics in Neonates during Sleep. *Entropy (Basel)* **19**, doi:10.3390/e19050225 (2017).
44 Ebrahimi, F., Setarehdan, S.-K. & Nazeran, H. Automatic sleep staging by simultaneous analysis of ECG and respiratory signals in long epochs. *Biomedical Signal Processing and Control* **18**, 69-79, doi:https://doi.org/10.1016/j.bspc.2014.12.003 (2015).
45 Malik, M. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Circulation* **93**, 1043-1065 (1996).
46 Cowan, M. J. Measurement of Heart Rate Variability. *Western Journal of Nursing Research* **17**, 32-48, doi:10.1177/019394599501700104 (1995).
47 Billman, G. Heart Rate Variability – A Historical Perspective. *Frontiers in Physiology* **2**, doi:10.3389/fphys.2011.00086 (2011).
48 Shaffer, F. & Ginsberg, J. P. An Overview of Heart Rate Variability Metrics and Norms. *Front Public Health* **5**, 258, doi:10.3389/fpubh.2017.00258 (2017).
49 P. Gomes, P. M., and H. P. da Silva. pyHRV: Development and evaluation of an open-source python toolbox for heart rate variability (HRV). *Proc. Int'l Conf. on Electrical, Electronic and Computing Engineering (IcETRAN)* (2019).
50 Welch, P. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics* **15**, 70-73, doi:10.1109/TAU.1967.1161901 (1967).
51 Fonseca, D. S., Netto, A. D., Ferreira, R. B. & Sá, A. M. F. L. M. d. in *2013 ISSNIP Biosignals and Biorobotics Conference: Biosignals and Robotics for Better and Safer Living (BRC)*. 1-4.
52 Miranda Dantas, E. *et al.* Spectral analysis of heart rate variability with the autoregressive method: What model order to choose? *Computers in Biology and Medicine* **42**, 164-170, doi:https://doi.org/10.1016/j.compbiomed.2011.11.004 (2012).
53 Estévez, M. *et al.* Spectral analysis of heart rate variability. *International Journal on Disability and Human Development* **15**, 5-17, doi:doi:10.1515/ijdhd-2014-0025 (2016).
54 Electrophysiology, T. F. o. t. E. S. o. C. t. N. A. S. o. P. Heart Rate Variability. *Circulation* **93**, 1043-1065, doi:doi:10.1161/01.CIR.93.5.1043 (1996).
55 Groth, D., Hartmann, S., Klie, S. & Selbig, J. in *Computational Toxicology: Volume II* (eds Brad Reisfeld & Arthur N. Mayeno) 527-547 (Humana Press, 2013).
56 Maleki, F. *et al.* Overview of Machine Learning Part 1: Fundamentals and Classic Approaches. *Neuroimaging Clinics of North America* **30**, e17-e32, doi:https://doi.org/10.1016/j.nic.2020.08.007 (2020).
57 Abdi, H. & Williams, L. J. Principal component analysis. *WIREs Computational Statistics* **2**, 433-459, doi:https://doi.org/10.1002/wics.101 (2010).
58 Ringnér, M. What is principal component analysis? *Nature Biotechnology* **26**, 303-304, doi:10.1038/nbt0308-303 (2008).
59 Harper, P. R. A review and comparison of classification algorithms for medical decision making. *Health Policy* **71**, 315-331, doi:https://doi.org/10.1016/j.healthpol.2004.05.002 (2005).
60 Müller, A. C. & Guido, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. (O'Reilly Media, Incorporated, 2016).
61 Zhang, C., Liu, C., Zhang, X. & Almpanidis, G. An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications* **82**, 128-150, doi:https://doi.org/10.1016/j.eswa.2017.04.003 (2017).
62 Fabian Pedregosa, G. V., Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830 (2011).

63 de Figueiredo, M. *et al.* A variable selection method for multiclass classification problems using two-class ROC analysis. *Chemometrics and Intelligent Laboratory Systems* **177**, 35-46, doi:https://doi.org/10.1016/j.chemolab.2018.04.005 (2018).

64 Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **27**, 861-874, doi:https://doi.org/10.1016/j.patrec.2005.10.010 (2006).

65 Bernard, S., Chatelain, C., Adam, S. & Sabourin, R. The Multiclass ROC Front method for cost-sensitive classification. *Pattern Recognition* **52**, 46-60, doi:https://doi.org/10.1016/j.patcog.2015.10.010 (2016).

66 Grandini, M., Bagli, E. & Visani, G. *Metrics for Multi-Class Classification: an Overview*. (2020).

67 Ambrogio, C., Koebnick, J., Quan, S. F., Ranieri, M. & Parthasarathy, S. Assessment of sleep in ventilator-supported critically III patients. *Sleep* **31**, 1559-1568, doi:10.1093/sleep/31.11.1559 (2008).

68 Fiorillo, L. *et al.* Automated sleep scoring: A review of the latest approaches. *Sleep Medicine Reviews* **48**, 101204, doi:https://doi.org/10.1016/j.smrv.2019.07.007 (2019).

69 Sun, H. *et al.* Sleep staging from electrocardiography and respiration with deep learning. *Sleep* **43**, zsz306, doi:10.1093/sleep/zsz306 (2020).

70 Vallat, R. & Walker, M. P. An open-source, high-performance tool for automated sleep staging. *eLife* **10**, e70092, doi:10.7554/eLife.70092 (2021).

71 Berry, R., Quan, S. and Abreu, A. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, version 2.6. *American Academy of Sleep Medicine,* (2020).

72 Bonnet, M. H. & Arand, D. L. Heart rate variability: sleep stage, time of night, and arousal influences. *Electroencephalogr Clin Neurophysiol* **102**, 390-396, doi:10.1016/s0921-884x(96)96070-1 (1997).

73 Somers, V. K., Dyken, M. E., Mark, A. L. & Abboud, F. M. Sympathetic-nerve activity during sleep in normal subjects. *N Engl J Med* **328**, 303-307, doi:10.1056/nejm199302043280502 (1993).

74 Tobaldini, E. *et al.* Heart rate variability in normal and pathological sleep. *Frontiers in Physiology* **4**, doi:10.3389/fphys.2013.00294 (2013).

75 Perry, M. A. & Kudchadkar, S. R. in *Sleep in Critical Illness: Physiology, Assessment, and Its Importance to ICU Care* (eds Gerald L. Weinhouse & John W. Devlin) 273-289 (Springer International Publishing, 2022).

76 Elliott, R., McKinley, S. & Cistulli, P. The quality and duration of sleep in the intensive care setting: An integrative review. *International Journal of Nursing Studies* **48**, 384-400, doi:https://doi.org/10.1016/j.ijnurstu.2010.11.006 (2011).

77 Marsillio, L. E., Manghi, T., Carroll, M. S., Balmert, L. C. & Wainwright, M. S. Heart rate variability as a marker of recovery from critical illness in children. *PLOS ONE* **14**, e0215930, doi:10.1371/journal.pone.0215930 (2019).

78 Johnston, B. W., Barrett-Jolley, R., Krige, A. & Welters, I. D. Heart rate variability: Measurement and emerging use in critical care medicine. *Journal of the Intensive Care Society* **21**, 148-157, doi:10.1177/1751143719853744 (2020).

79 Karmali, S. N., Sciusco, A., May, S. M. & Ackland, G. L. Heart rate variability in critical care medicine: a systematic review. *Intensive Care Medicine Experimental* **5**, 33, doi:10.1186/s40635-017-0146-1 (2017).

80 Ruopp, M. D., Perkins, N. J., Whitcomb, B. W. & Schisterman, E. F. Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biom J* **50**, 419-430, doi:10.1002/bimj.200710415 (2008).

81 Landgrebe, T. C. W. & Duin, R. P. W. Approximating the multiclass ROC by pairwise analysis. *Pattern Recognition Letters* **28**, 1747-1758, doi:https://doi.org/10.1016/j.patrec.2007.05.001 (2007).

1. Post-optimisation techniques

Two-class context

ROC-curve

The used machine learning models assign a probability score for each sample belonging to a particular class. This probability score is converted into a class label based on the decision threshold. The default threshold in binary classification is equal to 0.5. This means that a sample is predicted to belong in this class if the probability score for the positive class is greater than 0.5.

The ROC curve is obtained by changing the decision threshold over a wide range of values. For each decision threshold value, it is examined to which true positive rate (TPR) and to which false positive rate (FPR) this corresponds, resulting in a point on the ROC curve. An example of a ROC curve is provided in Figure S1.

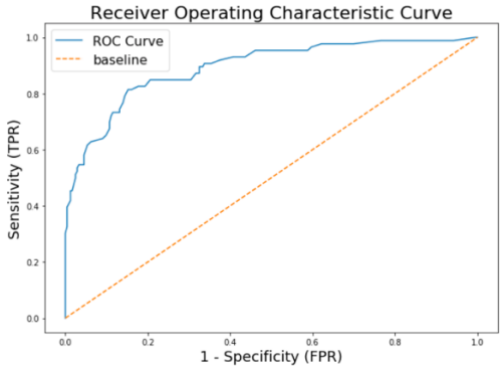


Figure S1. Example of a ROC curve.

Threshold moving

In certain cases, the default decision threshold is not an optimal representation of the predicted probabilities. In those cases, the ROC curve shows a good performance, where in reality the predictions do not match the actual classification well. This can be improved by optimising the threshold value.[65] A common method to determine the optimal threshold value is to search for the threshold resulting in maximum TPR and minimum FPR. This is obtained by maximising the Youden Index (TPR-FPR).[80]

Multi-class context

Optimising the threshold for a multi-class problem is much more complicated. For a problem with K classes, the used models per sample (x) produce a vector containing probabilities $P(\lambda_i|x)$, representing the probability that x belongs to class λ_i, with i from 1 to K:[65]

$$P(x) = [P(\lambda_1|x), P(\lambda_2|x), \dots, P(\lambda_K|x)]$$

For each sample, the probability scores assigned to the different classes sum up to one. An example is shown in Table S1.

Table S1. Example of the probability matrix for a three-class problem.

Sample	P(NREM)	P(REM)	P(Wake)
1	0.80	0.07	0.13
2	0.64	0.33	0.03
3	0.41	0.43	0.16

In multi-class problems, the class with the highest probability is assigned the class label for each sample. This highest probability class per sample can be obtained by taking the argmax over the probabilities:

$$\underset{i=1,..,K}{\operatorname{argmax}} P(\lambda_i|x) = argmax([P(\lambda_1|x), P(\lambda_2|x), \dots, P(\lambda_K|x)])$$

For a multi-class problem, ROC curves are created for each class separately by applying 'one versus the rest'. For a K-class problem, this leads to a total of K different ROC curves. To optimise the operating point on the ROC curves, several methods have been devised and investigated. Of these, the two best-performing methods have been further examined in both the diagnostic dataset using the XGBoost method (explanation shown in this subsection for the three-class problem) and in synthetic datasets using the logistic regression method (example shown in this subsection for a four-class problem).

Explanation of the post-optimisation methods applied on the diagnostic dataset

No post-optimisation technique applied

After applying the nCV procedure, the outer loop was run again using the optimal hyperparameters. For the XGBoost method with three classes this resulted in a balanced accuracy of 0.52. The corresponding confusion matrix and a hypnogram of one patient are shown in Figure S2 and Figure S3 respectively.

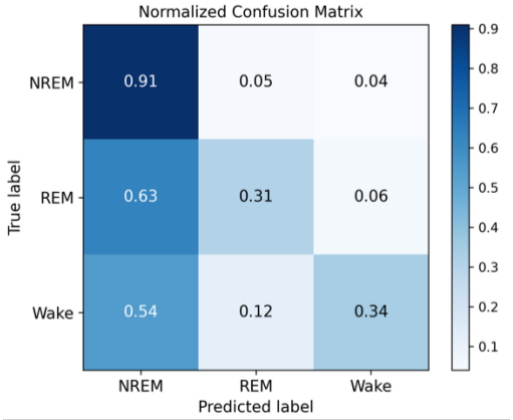


Figure S2. Confusion matrix of the observed stages versus the predicted stages calculated for the XGBoost classifier without applying post-optimisation techniques. The outer CV loop was used.

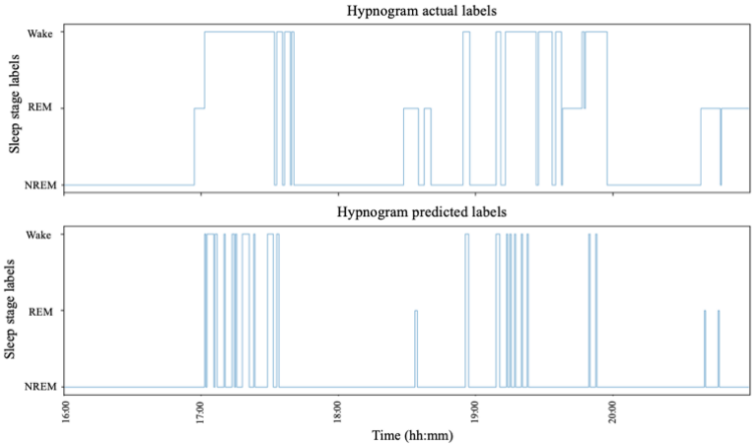


Figure S3. Hypnogram of one patient, showing the observed and predicted sleep stages over time, determined with the XGBoost classifier without applying post-optimisation techniques.

Method 1 – Optimisation of the individual thresholds

The first method optimises the threshold for each class K separately by considering the multi-class problem as K binary class problems: each class K versus the rest. An ROC curve is created for each class, where the optimal threshold per ROC curve is found by maximising the Youden Index. The optimal thresholds are subtracted from the original probabilities:

$$P(x) = [P(\lambda_1|x) - o_1, P(\lambda_2|x) - o_2, \dots, P(\lambda_K|x) - o_K]$$

The prediction then follows by extracting the maximum of the new probabilities:

$$argmax([P(\lambda_1|x) - o_1, P(\lambda_2|x) - o_2, \dots, P(\lambda_K|x) - o_K])$$

This method was applied to the posterior probabilities of the XGBoost method for three classes and resulted in a balanced accuracy over the five folds of 0.61. The corresponding confusion matrix and hypnogram are shown in Figure S4 and Figure S5 respectively.

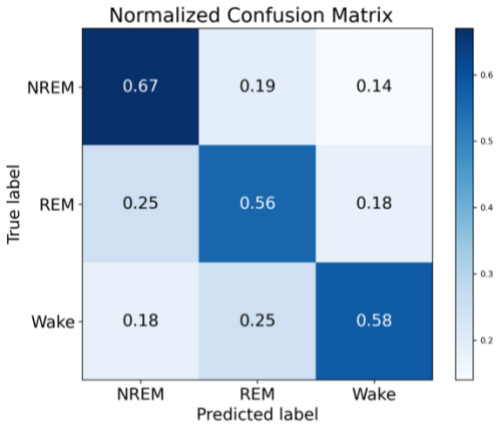


Figure S4. Confusion matrix of the observed stages versus the predicted stages calculated for the XGBoost classifier applying separate thresholds to the posterior probabilities. The outer CV loop was used.

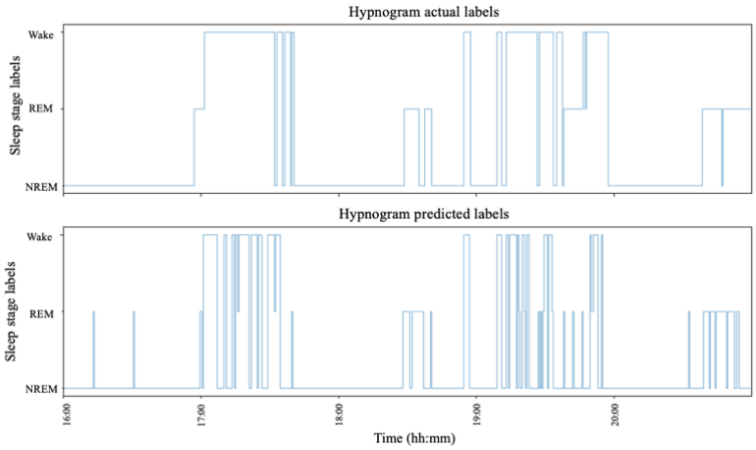


Figure S5. Hypnogram of one patient, showing the observed and predicted sleep stages over time, determined with the XGBoost applying separate thresholds to the posterior probabilities.

Method 2 – Using a weight vector

Each class probability is weighted with a scalar ω_i.[81] This means that all probabilities are multiplied by the weight vector $\omega = [\omega_1, \omega_2, \omega_3, \dots, \omega_k]$. Then the prediction follows from taking the argmax of the weighted probabilities:

$$argmax([\omega_1 \cdot P(\lambda_1|x), \omega_2 \cdot P(\lambda_2|x), \dots, \omega_K \cdot P(\lambda_K|x)])$$

To determine the optimal weights, the first weight was set to one. For the other weights, all combinations were considered with $10^{-2} \leq \omega_i \leq 10$ using a logarithmic scale. Finally, the combination of weights resulting in the highest balanced accuracy was used.

This weighting method was also applied to the posterior probabilities of the XGBoost model for three classes, resulting in a balanced accuracy of 0.61. The corresponding confusion matrix and hypnogram are shown in Figure S6 and Figure S7 respectively.

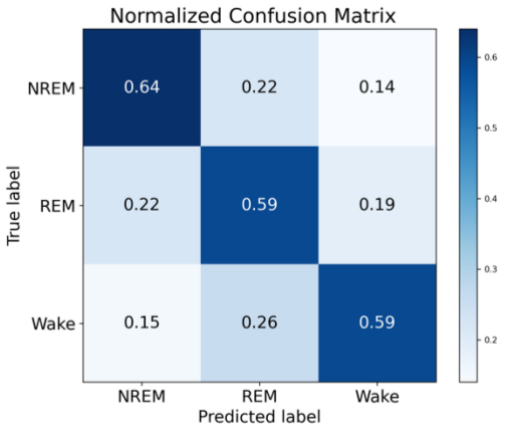


Figure S6. Confusion matrix of the observed stages versus the predicted stages calculated for the XGBoost classifier applying weighting to the posterior probabilities. The outer CV loop was used.

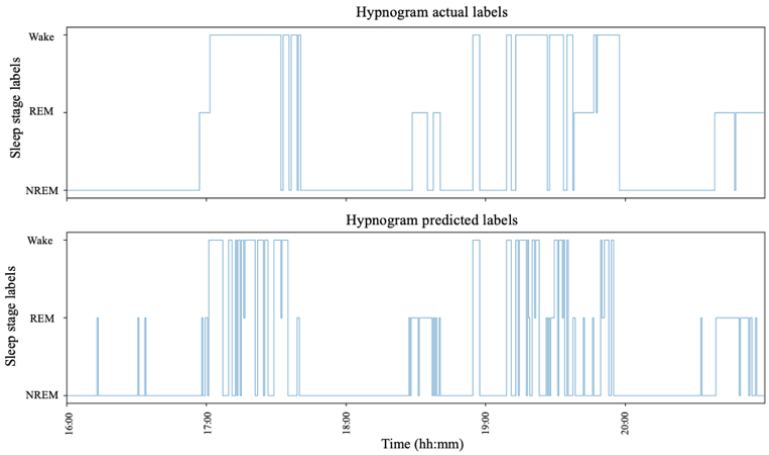


Figure S7. Hypnogram of one patient, showing the observed and predicted sleep stages over time, determined with the XGBoost applying weighting to the posterior probabilities.

Other methods

The above sections describe two methods to obtain optimal performance. In addition, two other methods have been considered. One of these methods was based on the order of the sleep stages. In a three-class problem, this order was: NREM – REM – Wake. First optimising NREM relative to the rest and then optimising Wake relative to the rest enabled the possibility to determine whether the samples were classified as NREM or wake. The unclassified remainder of the samples were then classified into REM. However, this method resulted in lower balanced accuracy than the individual threshold optimisation and weighted methods.

Post-optimisation experiments synthetic dataset

In addition to applying the post-optimisation methods to the diagnostic dataset, experiments were also performed on synthetic datasets, to verify the applicability and efficacy of the methods. Highlighted is a synthetic problem generated with the sklearn toolbox, consisting of four classes with a skewed distribution (group 1: 1.5%, group 2: 54.0%, group 3: 7.5%, group 4: 38.0%). In this experiment, the logistic regression algorithm was trained and optimised on 2700 samples and tested on 300 samples. The resulting default decision boundary on the test set is shown in Figure S8. The current decision boundaries predict the majority classes (group 2 and group 4) particularly well. Group 3 is predicted correctly in 70% of the cases, while group 1 is never predicted correctly. This leads to a balanced accuracy of 0.61.

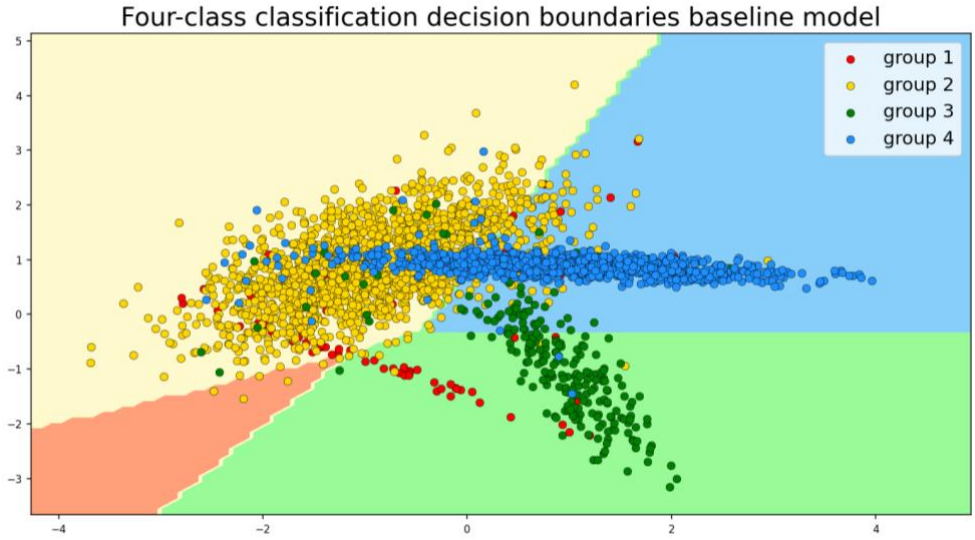


Figure S8. Default decision boundaries for a four-class problem with a skewed class distribution.

Method 1 – optimisation of the individual thresholds

After applying Method 1 (Figure S9), the decision boundaries are shifted such that part of the samples belonging to group 1 are predicted as group 1. In addition, the decision boundary separating group 3 and group 4 has shifted such that samples belonging to group 3 are now largely predicted correctly. Shifting the decision boundaries to better predict the minority classes results in a slight decrease in correctly predicted samples in the majority classes. However, this leads to an improved balanced accuracy of 0.75.

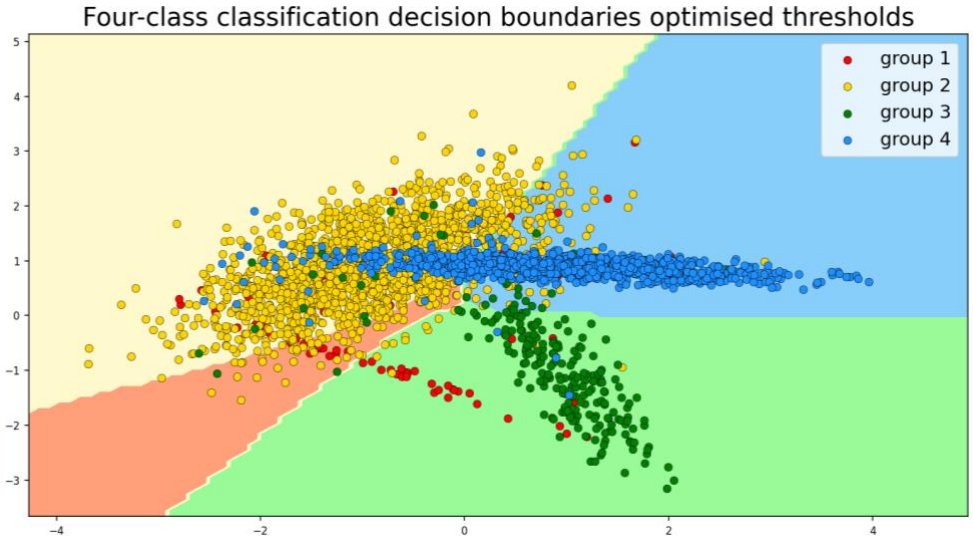


Figure S9. Decision boundaries for a four-class problem with a skewed class distribution optimised using the 'optimisation of individual thresholds' method.

Method 2 – Using a weight vector

Applying Method 2 (Figure S10) especially improves classification of the minority class, group 1, whose samples are now predicted well in 50% of the cases. This leads to an improved balanced accuracy of 0.78.

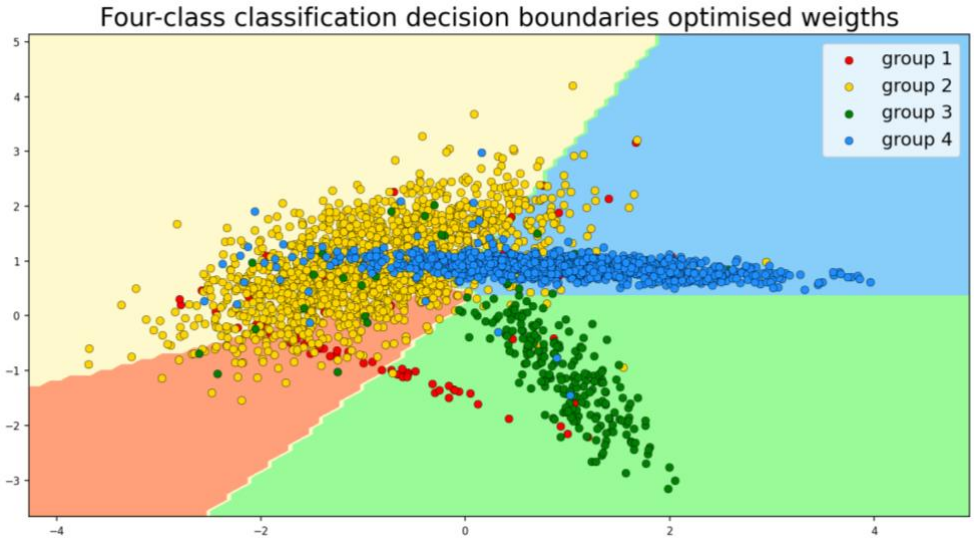


Figure S10. Decision boundaries for a four-class problem with a skewed class distribution optimised using a weight vector.

Insights obtained from experiments on synthetic datasets

The balanced accuracy improved in all studied synthetic datasets with the application of the optimisation techniques. However, the increase in balanced accuracy was minimal when the datasets were evenly distributed. In datasets with a skewed distribution, a larger increase was seen in balanced accuracy with both optimisation methods achieving comparable performance, with an often slightly higher performance in Method 2. The difference between the two methods increased in a more extreme situation, i.e., if one of the groups possesses less than 2% of the samples. This is because with Method 1, the minority class was hardly detected, if at all.

2. Dimensionality reduction

PCA can be applied to reduce the dimensionality of the data while preserving most of the information and variance within the dataset. Figure S11 presents the number of principal components required to explain a certain fraction of variance. Only 25 principal components are required to explain 95% of the variance. To explain 99% of the variance, 40 principal components are used.

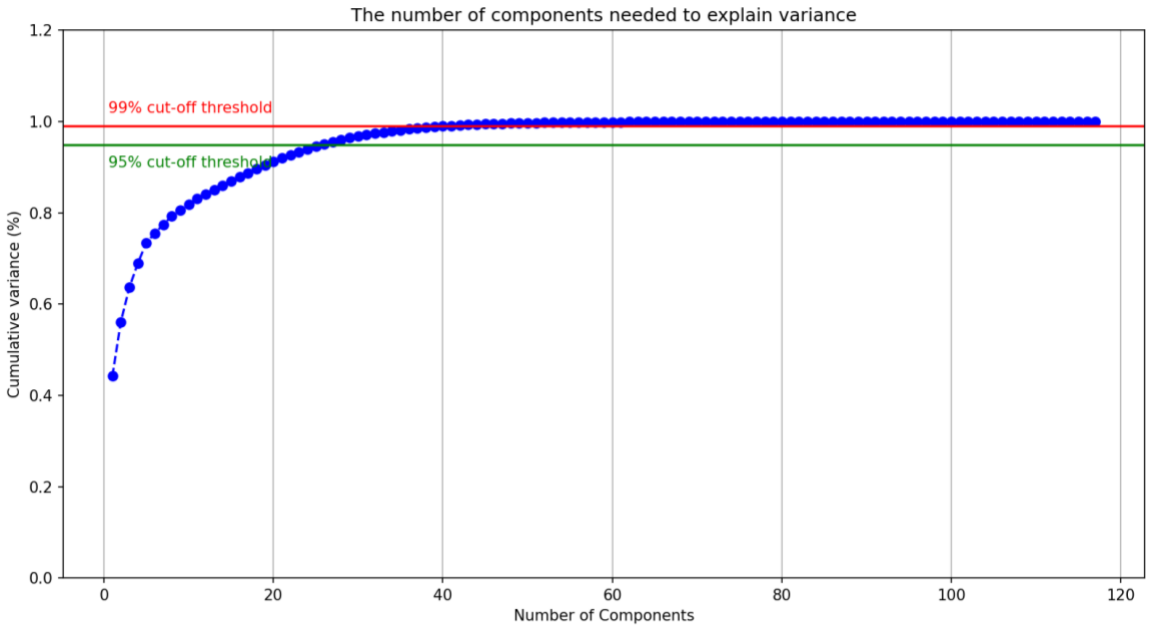


Figure S11. The number of principal components to explain variance.

3. Hyperparameter grid

Table S2 and Table S3 present respectively the hyperparameters included in the parameter grid during the nCV procedure of the models developed based on both ECG and PTT features and on ECG features only respectively. Other hyperparameters are set to their default values.

Table S2. Hyperparameter settings, parameter grid and final tuning configuration for the models developed based on both ECG and PTT features.

Hyperparameter per model	Settings baseline performance	Parameter grid	Final configuration 2-class staging	Final configuration 3-class staging	Final configuration 4-class staging	Final configuration 5-class staging
Logistic regression						
Solver	lbfgs	lbfgs	lbfgs	lbfgs	lbfgs	lbfgs
Penalty	L2	L2	L2	L2	L2	L2
C	1.0	0.001, 0.01, 0.1, 1.0, 10, 100	0.001	0.001	0.001	0.001
Class weight	None	balanced	balanced	balanced	balanced	balanced
Random forest						
Number of estimators	100	120, 300, 500 800	500	800	500	800
Max depth	6	5, 8	8	8	8	8
Min samples split	2	2, 5	5	5	5	5
Class weight	None	balanced	balanced	balanced	balanced	balanced
XGBoost						
Learning rate	0.3	0.05, 0.1, 0.2	0.1	0.1	0.1	0.1
Max depth	6	5, 8	5	8	5	5
Min child weight	1	1, 3	3	3	1	1
Subsample	1	0.8	0.8	0.8	0.8	0.8
Colsample by tree	1	0.8	0.8	0.8	0.8	0.8
Feature reduction						
PCA variance	0.95	0.99	0.99	0.99	0.99	0.99

Table S3. Hyperparameter settings, parameter grid and final tuning configuration for the models developed based on ECG features only.

Hyperparameter per model	Settings baseline performance	Parameter grid	Final configuration 2-class staging	Final configuration 3-class staging	Final configuration 4-class staging	Final configuration 5-class staging
Logistic regression						
Solver	lbfgs	lbfgs	lbfgs	lbfgs	lbfgs	lbfgs
Penalty	L2	L2	L2	L2	L2	L2
C	1.0	0.001, 0.01, 0.1, 1.0, 10, 100	0.001	0.001	0.001	0.001
Class weight	None	balanced	balanced	balanced	balanced	balanced
Random forest						
Number of estimators	100	120, 300, 500 800	300	800	800	800
Max depth	6	5, 8	8	8	8	8
Min samples split	2	2, 5	2	2	5	2
Class weight	None	balanced	balanced	balanced	balanced	balanced
XGBoost						
Learning rate	0.3	0.05, 0.1, 0.2	0.05	0.1	0.1	0.1
Max depth	6	5, 8	8	5	5	5
Min child weight	1	1, 3	3	3	1	3
Subsample	1	0.8	0.8	0.8	0.8	0.8
Colsample by tree	1	0.8	0.8	0.8	0.8	0.8
Feature reduction						
PCA variance	0.95	0.99	0.99	0.99	0.99	0.99

4. Baseline performance

Table S4 presents the baseline performance of the different models. The hyperparameters of the models are set to the values as shown in Table S2 and Table S3. Other hyperparameters are set to their default values.

Table S4. Overview of the baseline performance for all models using the five-fold cross-validation procedure on the diagnostic dataset.

	Diagnostic dataset					
	ECG and PTT features			**ECG features**		
	LR	RF	XGB	LR	RF	XGB
Two stages						
Accuracy	0.87 (0.01)	0.87 (0.01)	0.87 (0.01)	0.86 (0.01)	0.86 (0.01)	0.86 (0.02)
AUROC	0.78 (0.04)	0.78 (0.04)	0.80 (0.02)	0.76 (0.02)	0.77 (0.04)	0.77 (0.03)
Cohen's kappa	0.25 (0.06)	0.11 (0.08)	0.33 (0.06)	0.19 (0.05)	0.08 (0.06)	0.29 (0.06)
F1 score	0.30 (0.06)	0.12 (0.01)	0.40 (0.07)	0.23 (0.06)	0.09 (0.06)	0.36 (0.07)
Balanced accuracy	0.59 (0.02)	0.53 (0.03)	0.64 (0.03)	0.56 (0.02)	0.52 (0.02)	0.62 (0.03)
Three stages						
Accuracy	0.69 (0.02)	0.69 (0.02)	0.70 (0.02)	0.69 (0.02)	0.68 (0.02)	0.70 (0.02)
AUROC	0.76 (0.02)	0.76 (0.01)	0.78 (0.02)	0.76 (0.02)	0.77 (0.02)	0.78 (0.02)
Cohen's kappa	0.23 (0.04)	0.05 (0.03)	0.32 (0.03)	0.23 (0.04)	0.03 (0.02)	0.32 (0.03)
F1 score	0.47 (0.03)	0.32 (0.02)	0.54 (0.02)	0.47 (0.03)	0.30 (0.01)	0.54 (0.02)
Balanced accuracy	0.46 (0.02)	0.36 (0.01)	0.52 (0.02)	0.45 (0.02)	0.35 (0.01)	0.52 (0.02)
Four stages						
Accuracy	0.49 (0.01)	0.46 (0.02)	0.50 (0.01)	0.49 (0.01)	0.46 (0.02)	0.49 (0.01)
AUROC	0.73 (0.01)	0.73 (0.01)	0.74 (0.01)	0.73 (0.01)	0.73 (0.01)	0.74 (0.01)
Cohen's kappa	0.27 (0.01)	0.18 (0.02)	0.29 (0.02)	0.26 (0.01)	0.18 (0.02)	0.28 (0.01)
F1 score	0.45 (0.01)	0.34 (0.02)	0.48 (0.01)	0.45 (0.01)	0.35 (0.02)	0.47 (0.01)
Balanced accuracy	0.45 (0.01)	0.36 (0.01)	0.48 (0.01)	0.44 (0.01)	0.36 (0.01)	0.47 (0.01)
Five stages						
Accuracy	0.44 (0.02)	0.43 (0.02)	0.43 (0.03)	0.44 (0.02)	0.43 (0.02)	0.43 (0.02)
AUROC	0.73 (0.01)	0.72 (0.02)	0.72 (0.02)	0.73 (0.01)	0.72 (0.02)	0.72 (0.02)
Cohen's kappa	0.25 (0.02)	0.22 (0.03)	0.25 (0.03)	0.25 (0.03)	0.22 (0.02)	0.25 (0.03)
F1 score	0.36 (0.02)	0.34 (0.02)	0.39 (0.02)	0.36 (0.02)	0.34 (0.02)	0.38 (0.02)
Balanced accuracy	0.38 (0.01)	0.35 (0.01)	0.39 (0.02)	0.38 (0.02)	0.35 (0.01)	0.30 (0.02)

All values are presented as mean (standard deviation) over the five folds. LR = logistic regression, RF = random forest, XGB = XGBoost

5. Internal validation age categories

To gain insight into potential differences in classification performance at different ages, the optimised models were trained on part of the training data and then tested on 10 patients (not included in the training part) for each age category. Results are presented in Table S5.

Table S5. Performance of the machine learning models developed based on ECG features for four-class staging per age category.

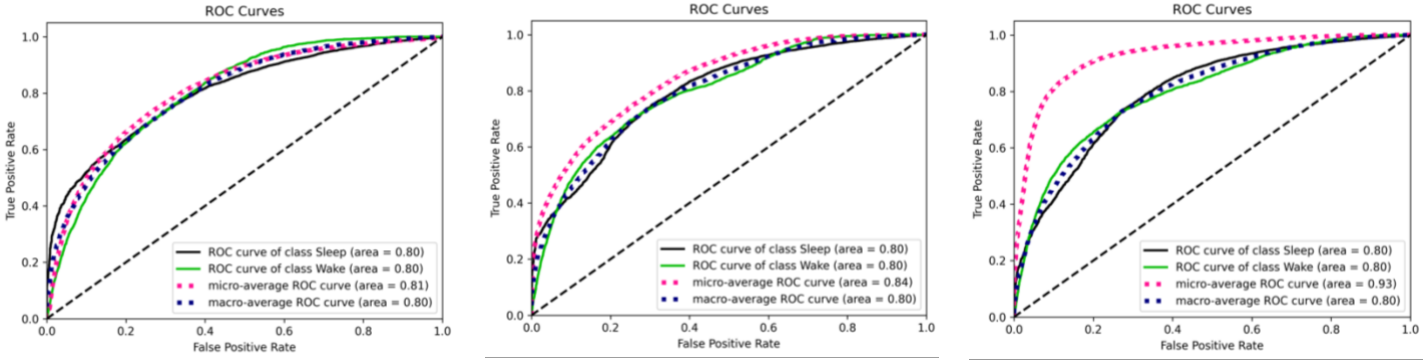
	AUROC, mean (SD)		
Age category	**LR**	**RF**	**XGB**
6-12 mos	0.78 (0.05)	0.78 (0.05)	0.79 (0.05)
1-3 y	0.78 (0.05)	0.76 (0.05)	0.78 (0.05)
3-5 y	0.78 (0.05)	0.76 (0.06)	0.77 (0.05)
5-9 y	0.73 (0.09)	0.73 (0.09)	0.73 (0.08)
9-13 y	0.76 (0.05)	0.73 (0.06)	0.72 (0.07)
13-18 y	0.77 (0.06)	0.76 (0.05)	0.77 (0.06)
p-value *	0.52	0.46	0.18

LR = logistic regression, RF = random forest, XGB = XGBoost.
** Calculated using one-way ANOVA.*

6. Nested cross-validation ROC curves and confusion matrices

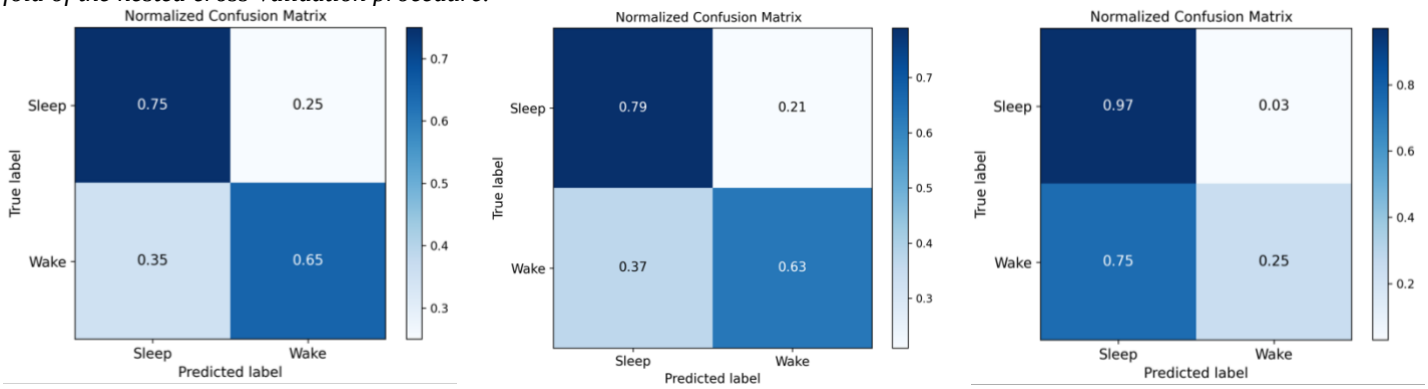
The ROC-curves for the models developed based on ECG features alone for two-class staging and five-class staging are presented in respectively Figure S12 and Figure S14. The associated confusion matrices are presented in Figure S13 and Figure S15.

The AUROC values decrease with an increase in the number of classes to be distinguished. The ROC curves presented in Figure S14 show that stages NREM 1 and NREM 2 have the lowest AUROC values. This indicates that these stages are the least distinctive. The associated confusion matrices (Figure S15) show that stages NREM 3, REM and wake are generally predicted correctly more than half of the time. NREM 1 and NREM 2, on the other hand, are mostly predicted incorrectly. Samples belonging to class NREM 1 are often misclassified as REM or wake. Samples belonging to class NREM 2 are mainly classified as NREM 3.



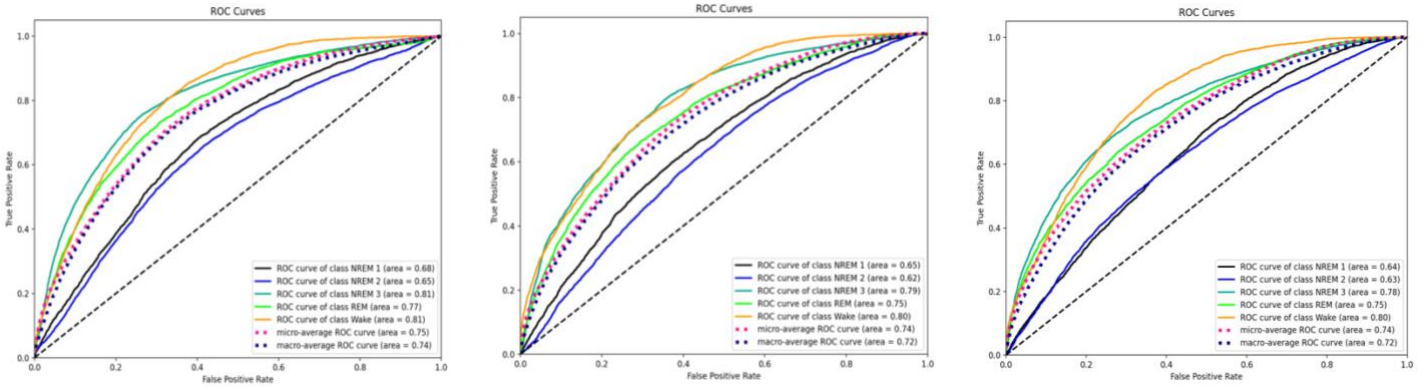
(a) Logistic regression *(b) Random forest* *(c) XGBoost*

Figure S12. ROC curves of the models developed based on ECG features only for two-class staging (sleep-wake) obtained in the last fold of the nested cross-validation procedure.



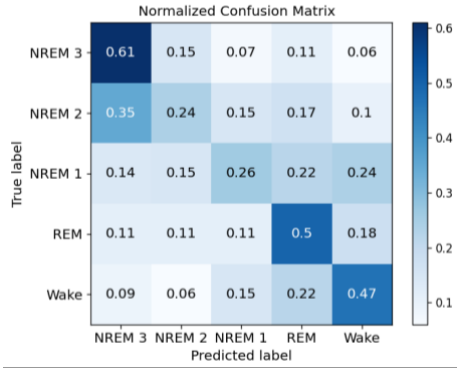
(a) Logistic regression *(b) Random forest* *(c) XGBoost*

Figure S13. Two-class staging confusion matrices of the actual stages versus the predicted stages (Sleep – wake). Calculated for all models developed based on ECG features only during the nested cross-validation procedure.

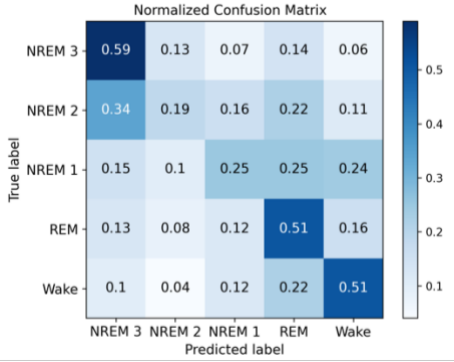


(a) Logistic regression *(b) Random forest* *(c) XGBoost*

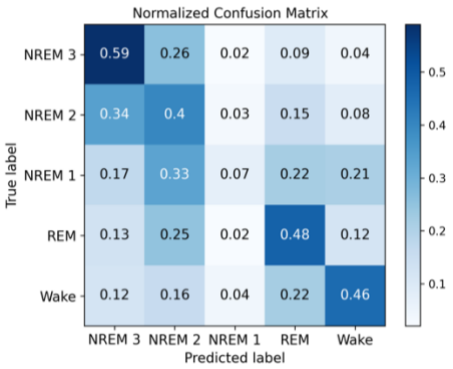
Figure S14. ROC curves of the models developed based on ECG features only for five-class staging (NREM 3 – NREM 2 – NREM 1 – REM – wake) obtained in the last fold of the nested cross-validation procedure.



(a) Logistic regression



(b) Random forest



(c) XGBoost

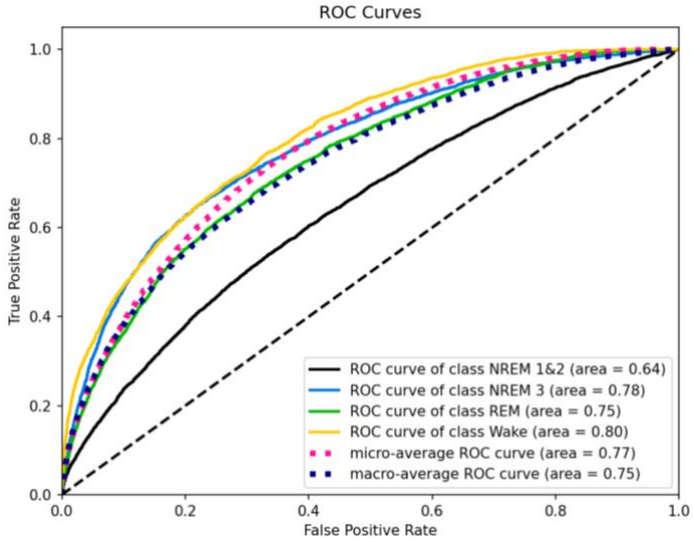
Figure S15. Five-class staging confusion matrices of the actual stages versus the predicted stages (NREM 3 – NREM 2 – NREM 1 – REM – wake). Calculated for all models developed based on ECG features only during the nested cross-validation procedure.

7. XGBoost performance

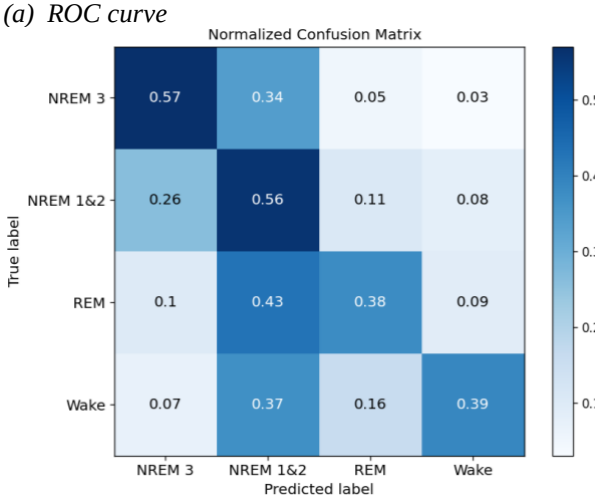
To gain insight into the additional performance measures of the classifiers, an overview of the results of the XGBoost model developed based on ECG features alone is presented here.

Figure S16 summarises the results of the XGBoost model for four-class staging. The ROC curve shows that stages NREM 3, REM and wake can be distinguished well from each other. However, NREM 1&2 shows to be less distinguishable with an AUROC value of 0.64. The confusion matrix obtained from the nested cross-validation procedure shows that most of the samples are classified as NREM 1&2 and contrary to what the ROC curve shows, REM and wake are often not properly classified. This is presumably because the thresholds have not been chosen properly.

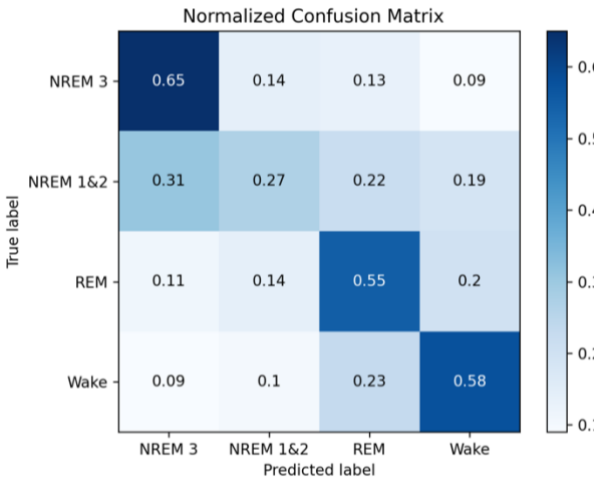
After applying the weight vector to the posterior probabilities, as explained in section 'Supplementary materials 1', the confusion matrix as shown in Figure S16c is obtained. The post-optimisation procedure ensured that most of the samples are no longer classified as NREM 1&2. Stages NREM 3, REM and wake are predicted with higher accuracies, contributing to the increase in balanced accuracy. The samples belonging to stage NREM 1&2 are predicted spread-across all stages, most often in NREM 1&2 and NREM 3 and least often in wake. It can be concluded that stage NREM 1&2 is difficult to distinguish from the other stages.



(a) ROC curve



(b) Confusion matrix nCV



(c) Confusion matrix after applying the post-optimisation procedure

Figure S16. Additional model performance analysis for the XGBoost classifier developed for four-class staging based on ECG features only. (a) shows the ROC curves with the AUROC values listed. (b) shows the confusion matrix obtained from the nested cross-validation procedure. (c) shows the confusion matrix after applying weights to the posterior probabilities.

Table S6 summarises the balanced accuracies before and after the post-optimisation procedure. The balanced accuracies are calculated by taking the average over the five folds. Table S7 provides an overview of the optimal weights as obtained from the post-optimisation procedure.

Table S6. Overview of the balanced accuracies after applying the nested cross-validation procedure to the diagnostic dataset and after optimisation of the posterior probabilities.

	ECG and PTT features		ECG features	
	After nCV	After post optimisation	After nCV	After post optimisation
Two stages	0.64 (0.04)	0.76 (0.02)	0.62 (0.04)	0.72 (0.02)
Three stages	0.54 (0.02)	0.62 (0.02)	0.52 (0.03)	0.61 (0.03)
Four stages	0.49 (0.01)	0.52 (0.01)	0.48 (0.01)	0.51 (0.01)
Five stages	0.41 (0.02)	0.42 (0.01)	0.40 (0.02)	0.41 (0.01)

Values are presented as mean (standard deviation).

Table S7. Weights obtained from the post-optimisation procedure.

	PTT and ECG features				ECG features			
	Two stages	Three stages	Four stages	Five stages	Two stages	Three stages	Four stages	Five stages
ω_1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
ω_2	6.4	3.8	2.4	3.0	7.2	3.8	1.6	2.8
ω_3	0.0	5.0	3.2	4.0	0.0	4.8	2.4	2.8
ω_4	0.0	0.0	3.8	5.6	0.0	0.0	3.2	3.8
ω_5	0.0	0.0	0.0	7.0	0.0	0.0	0.0	5.2

8. External validation on PICU data

The distribution of sleep stages varies greatly per patient admitted to the PICU. Figure S17 presents these differences compared to the distribution in the diagnostic dataset.

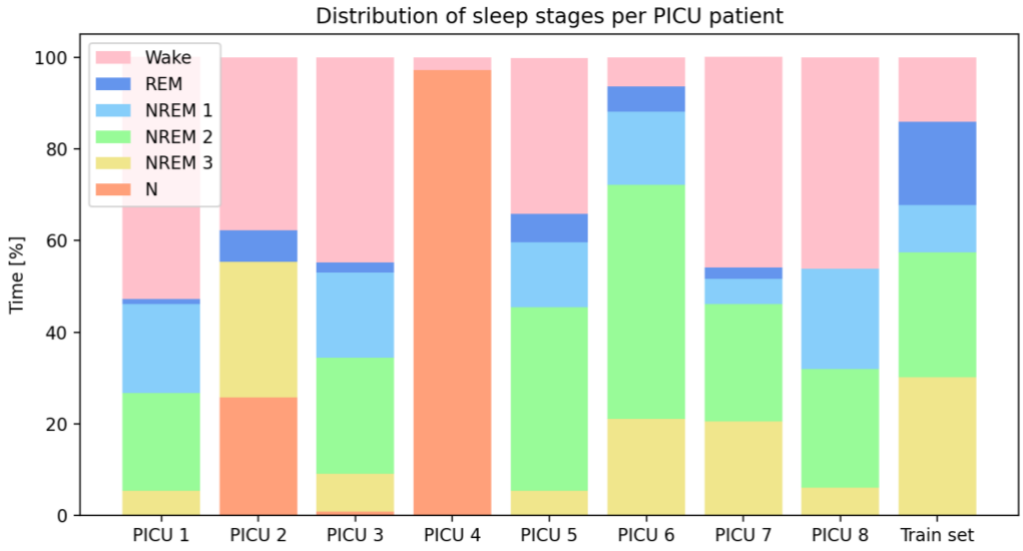
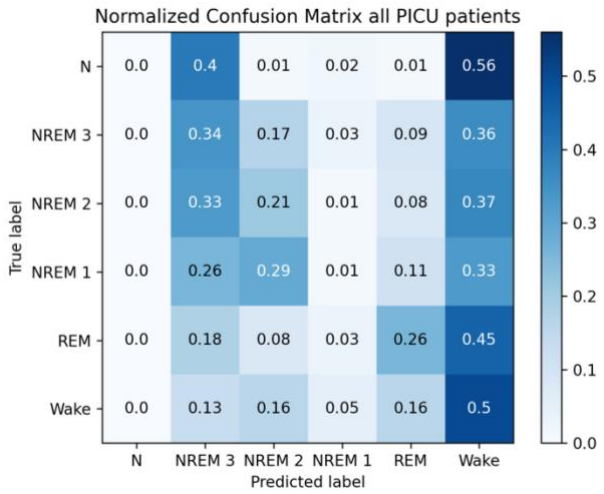
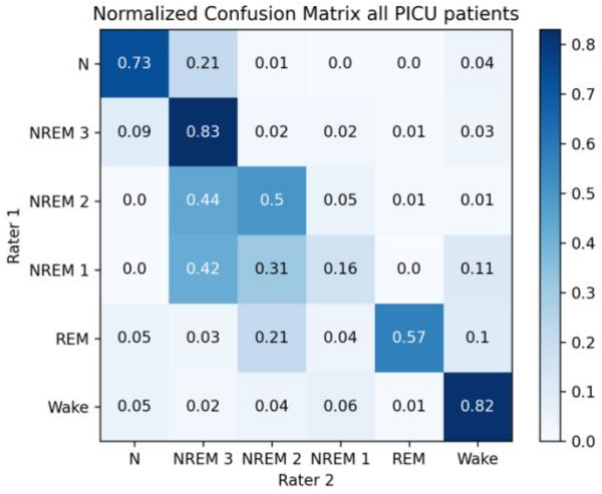


Figure S17. Distribution of sleep stages per PICU patient compared to the distribution of sleep stages in the diagnostic dataset.

For external validation, the final optimised models developed based on ECG features alone were fitted once more on the diagnostic dataset and then applied to the PICU dataset. The resulting confusion matrix of the actual stages with respect to the predicted stages for the random forest model developed for five-class staging is illustrated in Figure S18a. Figure S18b shows the confusion matrix of the two technicians with respect to each other.



(a) Random forest model



(b) Interrater agreement

Figure S18. Confusion matrices for all PICU patients.

9. Post optimisation external validation on PICU data

After external validation was completed, classification performance reduced strongly. To investigate whether the cause was a poorly calibrated model or whether it was due to a model that was not generalisable to the PICU dataset, an experiment was carried out in which the calibration of the models was improved using weights as described in 'Supplementary materials 1'. The balanced accuracies of the models before and after applying this post optimisation procedure are presented in Table S8 and Figure S19. Subsequently, the differences per PICU patient are also shown in Figure S20.

Table S8. Overview of the balanced accuracies obtained from external validation and after optimisation of the posterior probabilities.

	Logistic regression		Random forest		XGBoost	
	External validation	Post optimisation	External validation	Post optimisation	External validation	Post optimisation
Two stages	0.57 (0.45 – 0.69)	0.61 (0.51 – 0.72)	0.62 (0.46 – 0.75)	0.63 (0.47 – 0.75)	0.56 (0.42 – 0.70)	0.58 (0.49 – 0.69)
Three stages	0.43 (0.31 – 0.54)	0.51 (0.44 – 0.60)	0.47 (0.33 – 0.59)	0.50 (0.43-0.59)	0.46 (0.31 – 0.59)	0.49 (0.42 – 0.58)
Four stages	0.27 (0.17 – 0.34)	0.32 (0.26 – 0.37)	0.28 (0.19 – 0.37)	0.30 (0.24 – 0.35)	0.27 (0.17 – 0.35)	0.29 (0.24 – 0.34)

Values are illustrated with 95% CI values between brackets, obtained using bootstrapping with replacement across the patients 500 times.

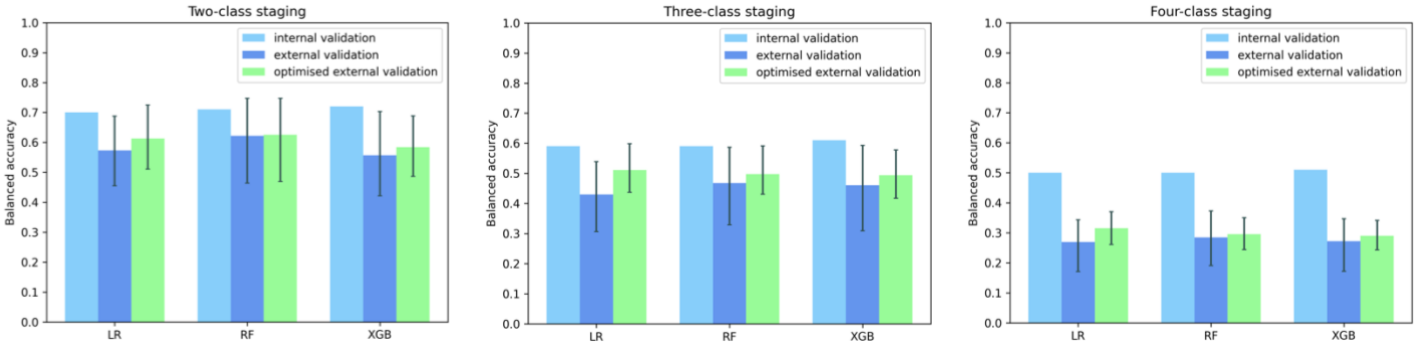


Figure S19. Balanced accuracies of internal validation, external validation and optimised external validation for all models. The confidence intervals for external validation obtained with bootstrapping are illustrated by the error bars.

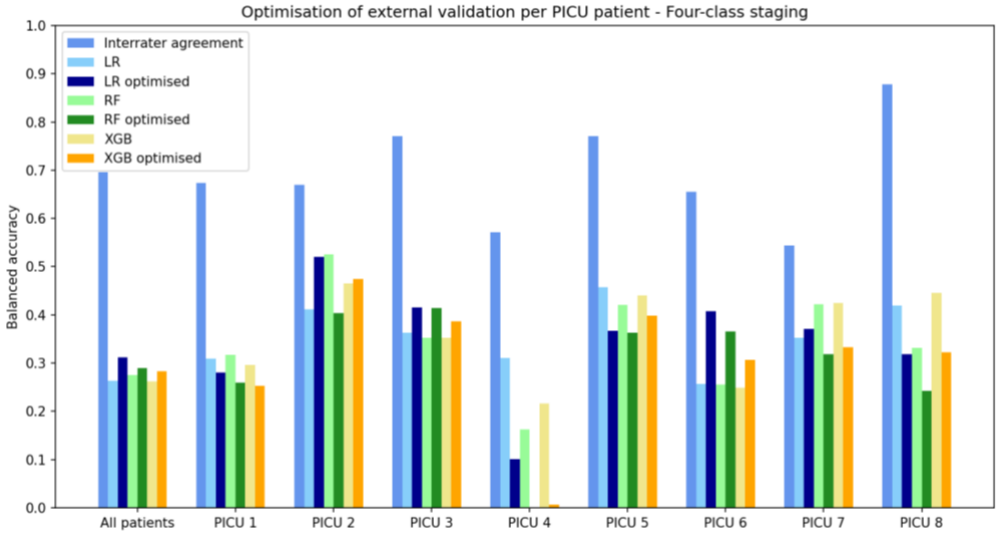


Figure S20. Balanced accuracies per PICU patient for all models for four-class staging. The balanced accuracies of external validation, optimised external validation and the interrater agreement are compared.

10. Hypnograms PICU patients

The hypnograms of the individual PICU patients obtained during external validation of the random forest model for four-class staging are presented in Figure S21 to Figure S28.

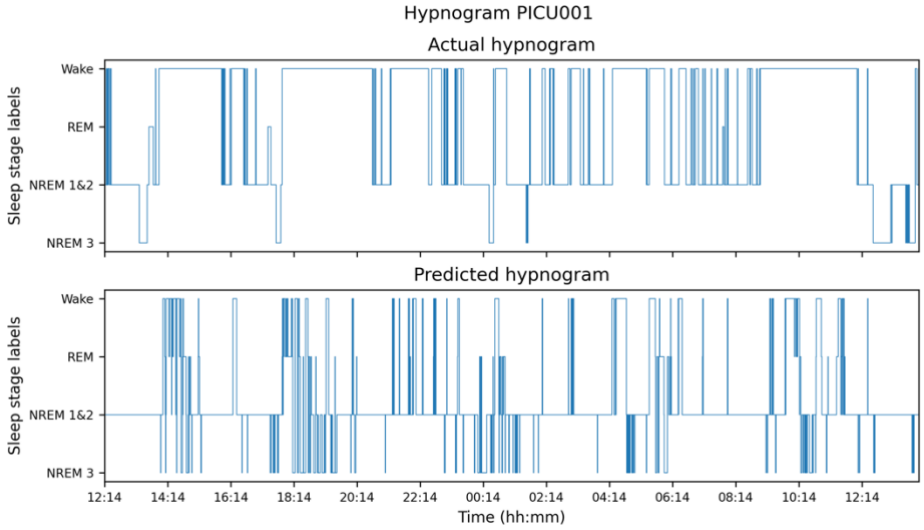


Figure S21. Hypnogram of PICU patient 1.

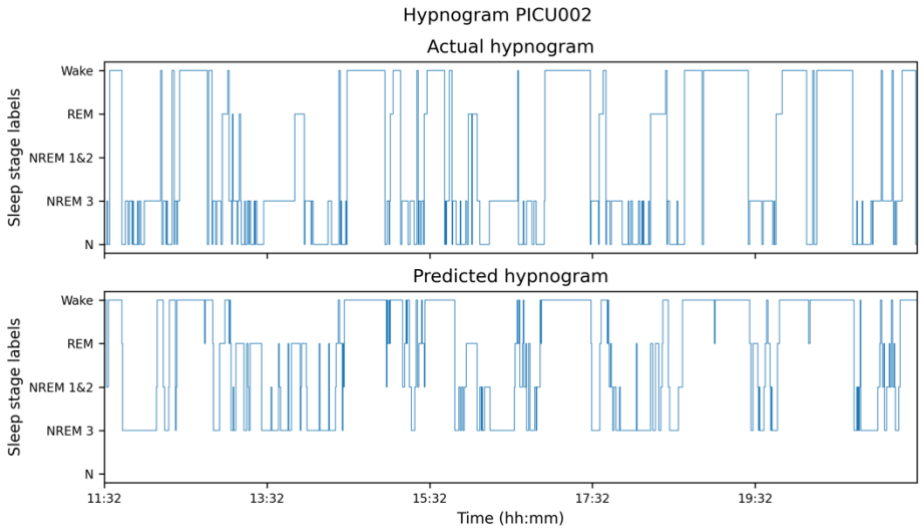


Figure S22. Hypnogram of PICU patient 2.

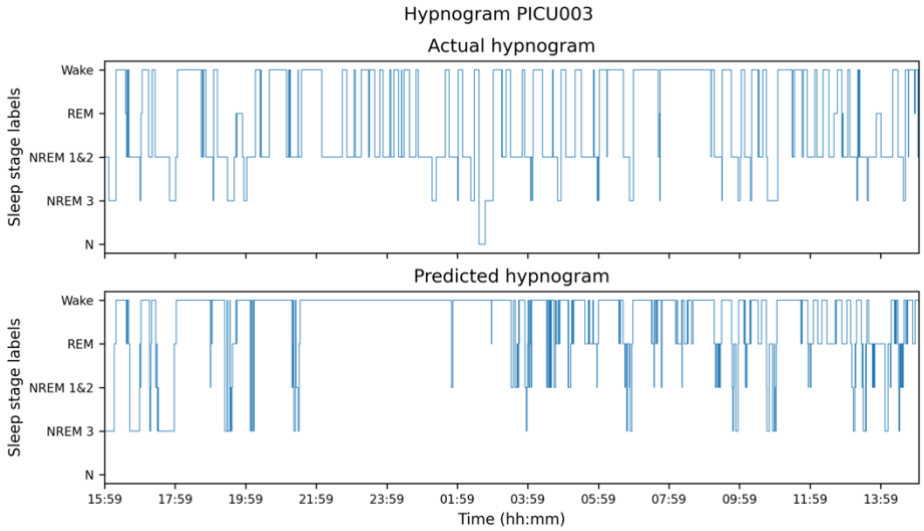


Figure S23. Hypnogram of PICU patient 3.

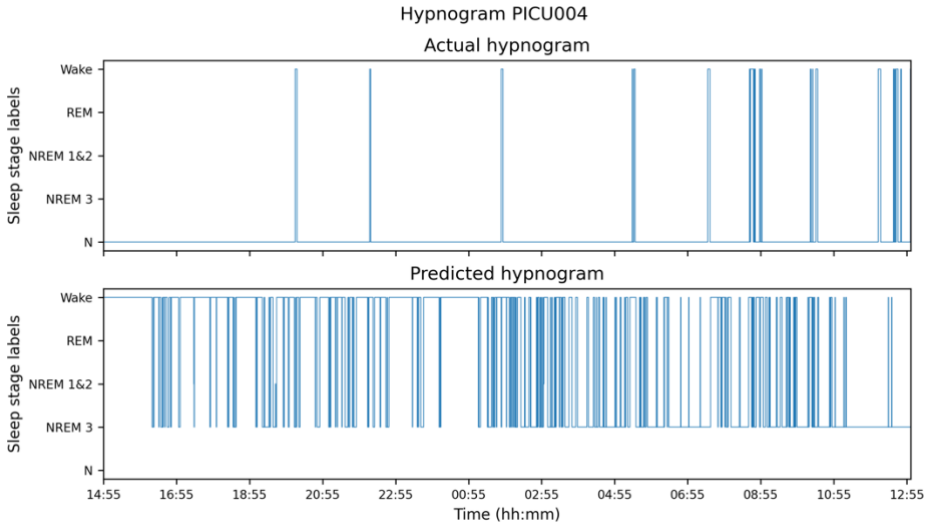


Figure S24. Hypnogram of PICU patient 4.

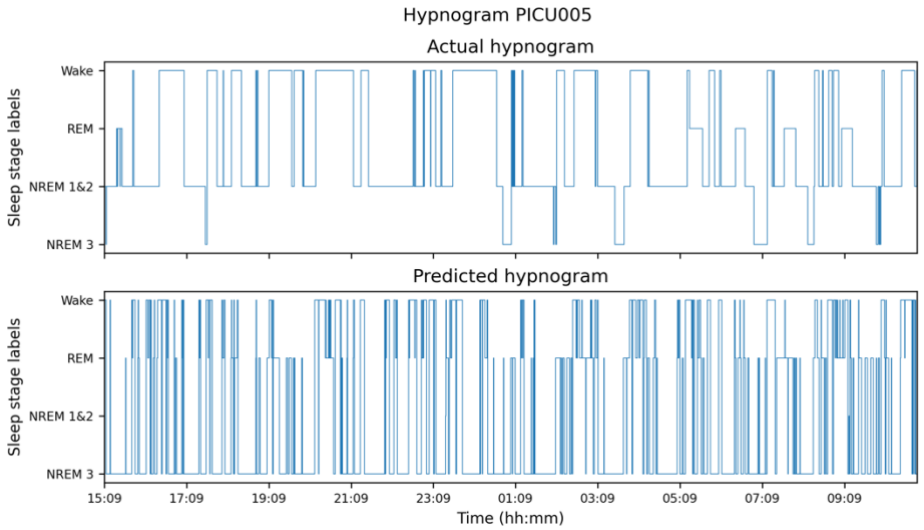


Figure S25. Hypnogram of PICU patient 5.

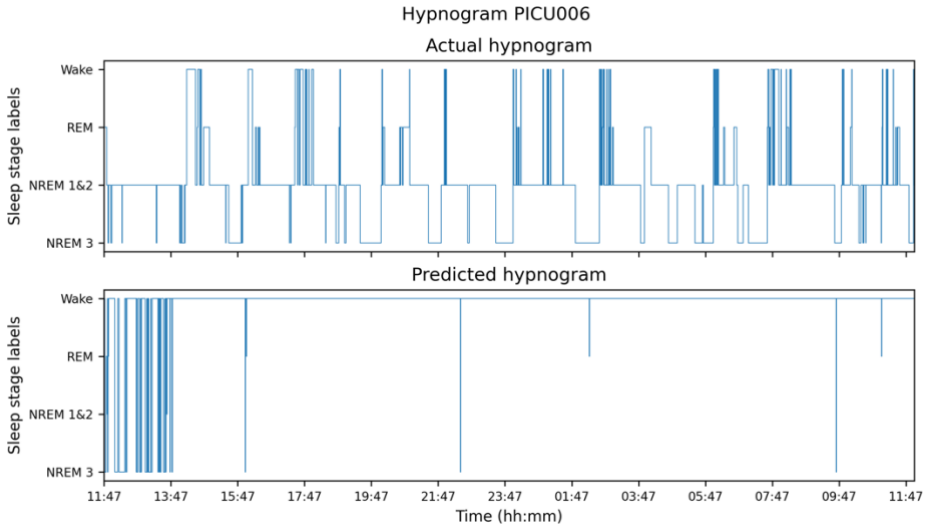


Figure S26. Hypnogram of PICU patient 6.

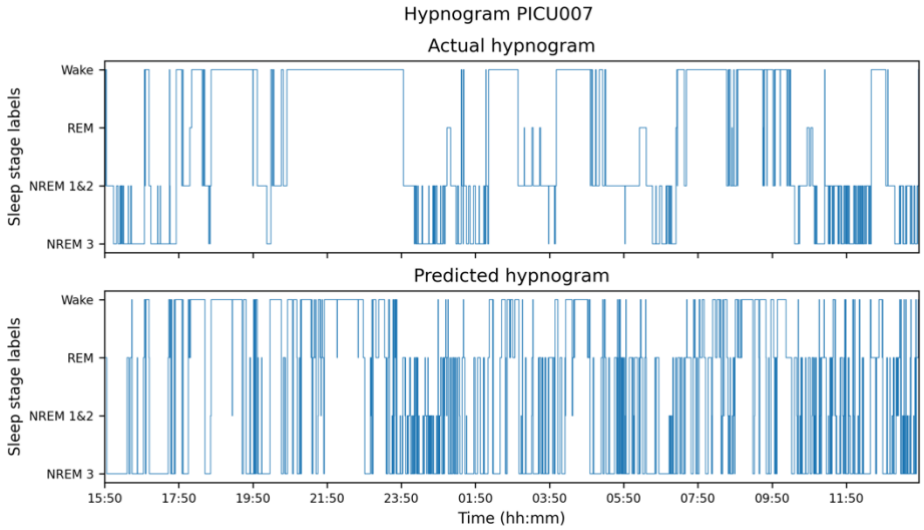


Figure S27. Hypnogram of PICU patient 7.

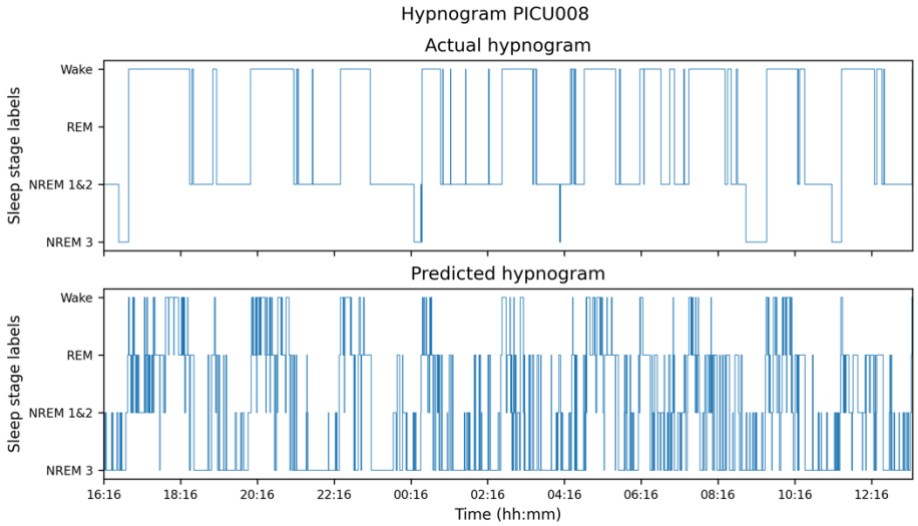


Figure S28. Hypnogram of PICU patient 8.