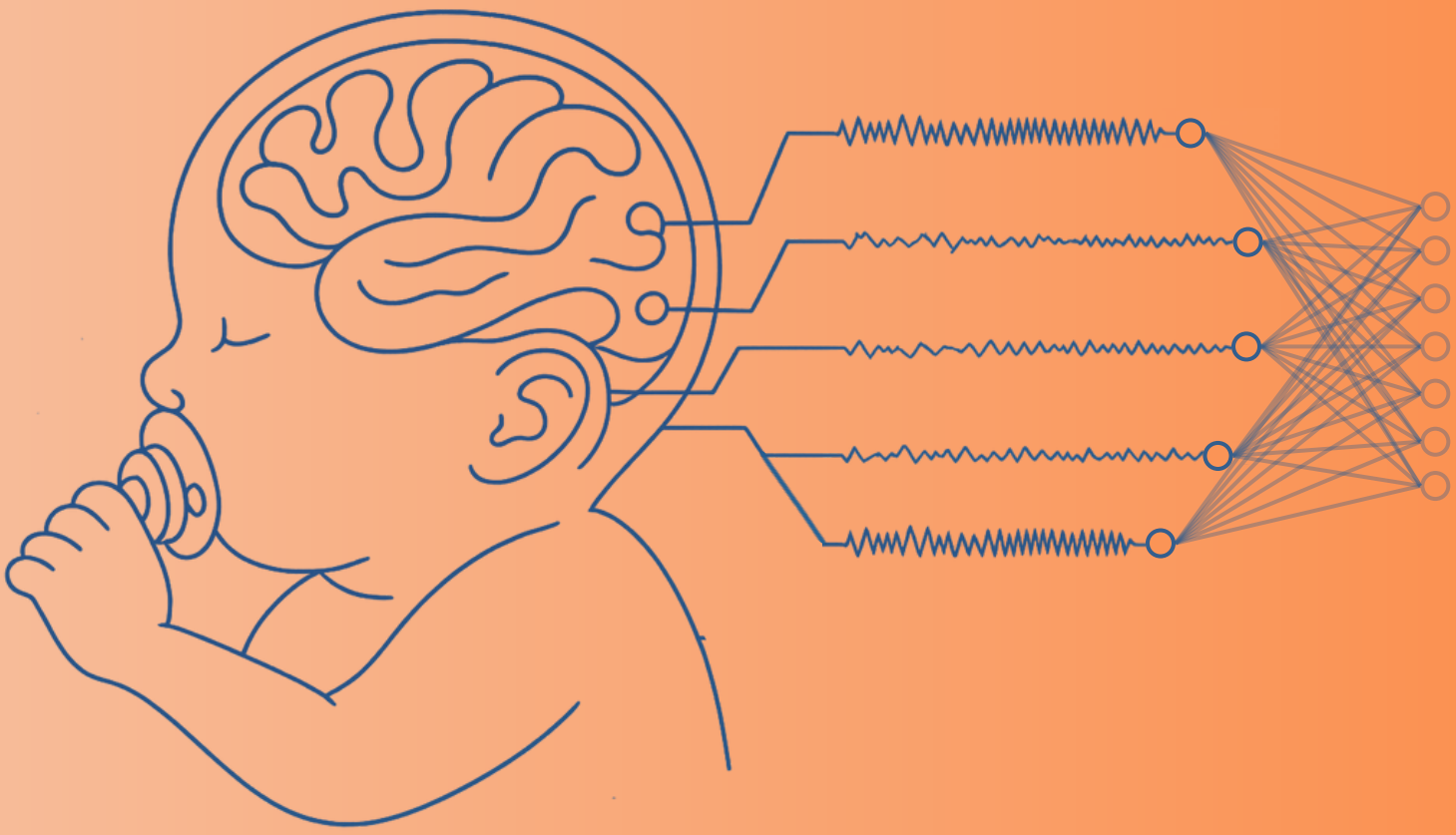
Supervised and Unsupervised Dynamic Machine Learning Approaches for Automatic EEG-Based Sleep Staging in Critically Ill Children



MARLOES JAGER

2025

MSc. Technical Medicine Thesis

Supervised and Unsupervised Dynamic Machine Learning Approaches for Automated EEG-Based Sleep Staging in Critically Ill Children

Marloes Jager
Student number: 4807960
25 April 2025

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 Pediatric Intensive Care Unit,
Erasmus MC Sophia Children's Hospital
28th Oktober 2024 – 25th April 2025

Supervisors:
dr. D.M.J. Tax, Assistant Professor, TU Delft
dr. R.C.J. de Jonge, MD, PhD, Erasmus MC
dr. J.W. Kuiper, MD, PhD, Erasmus MC
E. van Twist, MSc, Erasmus MC
B. van Winden, MSc, Erasmus MC

Thesis committee members:
dr. R.C.J. de Jonge, MD, PhD, Erasmus MC
dr. D.M.J. Tax, Assistant Professor, TU Delft
dr.ir. J.H. Krijthe, Assistant Professor, TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Universiteit Leiden TU Delft Delft University of Technology Erasmus Universiteit Rotterdam

Preface

With this thesis, my study period comes to an end. These past seven years have been a fun and eventful chapter of my life, both academically and beyond, starting in Delft and later in Rotterdam. I really enjoyed the study, especially the moments during internships when I got to explore what parts of technical medicine interest me most. Two major highlights worth mentioning were my trips to Tanzania and Nepal during my Master's, where I could experience healthcare in completely different settings and make many valuable memories.

I could not have done this thesis on my own, and therefore I would like to thank my supervisors. When choosing where to graduate, good supervision was a very important factor for me. After a TM2 internship at the Pediatric Intensive Care team at the Sophia Children's Hospital, I thought this would be the right department to complete my thesis, and I still believe this was the right choice.

Eris and Brian, thank you for literally always having the door open and making so much time for students. I really appreciated your openness, your willingness to brainstorm about anything, and the relaxed atmosphere you create together. Rogier and Jan Willem, thank you for your enthusiasm during our meetings, which always gave me positive energy. Even when the topic got a bit technical, your clinical and academic perspectives were always very helpful. And of course, David, thank you for your valuable time and energy. I always looked forward to our Monday meetings, which were always a motivating and positive start to the week. I enjoyed our interesting discussions, and they encouraged me to learn a lot on the topic of machine learning. Lastly, thank you to Jesse for being part of my thesis committee.

I'd also like to thank my (study) friends for all the study sessions and coffee breaks, they were much appreciated. And a special thanks to my housemates, who were all also working on their theses, making it valuable to share the process of writing a thesis together.

Marloes Jager
Rotterdam, April 2025

Abstract

Introduction: In critically ill paediatric patients, sleep is essential for recovery and development, yet sleep disturbances are common in the paediatric intensive care unit (PICU), which highlight the need to integrate sleep monitoring into clinical practice. While automated sleep stage classification using machine learning (ML) on single-channel electroencephalography (EEG) data has shown promise in mostly healthy adult populations, its application to critically ill children is challenged by age-specific sleep architecture, medication effects, pathological conditions, and artefacts. This study evaluates whether deep learning (DL) feature extraction and dynamic models can improve sleep staging performance in this population and explores the use of an unsupervised ML model to gain deeper insight into the complex sleep structures in this population.

Methods: This study utilised EEG recordings from three datasets—healthy adults, non-critically ill children, and critically ill children—to train and evaluate supervised and unsupervised sleep stage classification models. As supervised models, a convolutional neural network (CNN) was used for feature extraction, followed by dynamic models including a long short-term memory (LSTM) network and a hidden Markov model (HMM) to account for temporal dependencies in sleep. Additionally, an unsupervised HMM was applied to explore underlying structures in the sleep EEG data without predefined labels.

Results: Supervised models achieved good performance in healthy adults and non-critically ill children, with maximum accuracies of 90.2% and 77.4%, respectively, for three-state classification. The added value of dynamic models over the CNN alone varied per dataset and model type and was not consistent. In critically ill children, classification performance was low, with a maximum accuracy of 61.4%, and notably low macro-F1 and Cohen's kappa scores (45.9% and 26.5%, respectively). The unsupervised HMM revealed that identifying distinct and stable clusters over time was challenging in all datasets. For critically ill children, the model often failed to identify multiple distinct clusters within individual patients, and substantial variability in cluster assignments was observed across patients.

Discussion: This study demonstrates that DL-based feature extraction and dynamic modelling using single-channel EEG can achieve strong sleep staging performance in healthy adults and non-critically ill children, highlighting the potential for (semi-)automated scoring tools in more stable populations. In contrast, performance in critically ill children was notably lower, likely due to factors such as high variability in sleep architecture, signal artefacts, limited data quality, and the uncertain reliability of manually assigned labels. These results suggest that conventional sleep stages do not generalise well to this population, and a purely data-driven, unsupervised approach does not offer a viable alternative. Overall, the findings emphasise the need for a larger dataset of critically ill children, further evaluation of relevant patient and data characteristics, the inclusion of alternative signals such as electrocardiography, and greater focus on model interpretability.

The code is available at https://github.com/BrianvanWinden/tm3-slaap.

Contents

List of Abbreviations

AASM	American Academy of Sleep Medicine
ARI	Adjusted Rand Index
AUC ROC	Area Under the Receiver Operating Characteristic Curve
BiLSTM	Bidirectional Long Short-Term Memory
C3	Central Electrode (left hemisphere)
CNN	Convolutional Neural Network
CV	Cross-Validation
Cz	Central Electrode (midline)
DL	Deep Learning
ECG	Electrocardiography
EDF	European Data Format
EEG	Electroencephalography
EMG	Electromyography
EOG	Electrooculography
F3	Frontal Electrode (left hemisphere)
FPz	Frontal Pole Electrode
HMM	Hidden Markov Model
ICU	Intensive Care Unit
IQR	Interquartile Range
kappa	Cohen's Kappa
LSTM	Long Short-Term Memory
MEC	Medical Ethical Committee
MF1	Macro-averaged F1 Score
ML	Machine Learning
N1	Non-Rapid Eye Movement Sleep Stage 1
N2	Non-Rapid Eye Movement Sleep Stage 2
N3	Non-Rapid Eye Movement Sleep Stage 3
SD	Standard Deviation
NIV	Non-Invasive Ventilation
NREM	Non-Rapid Eye Movement Sleep
NSWS	Non-Slow-Wave Sleep
OSAS	Obstructive Sleep Apnoea Syndrome
PCA	Principal Component Analysis
PELOD-2	Paediatric Logistic Organ Dysfunction score, version 2
PICU	Paediatric Intensive Care Unit
PIM-3	Paediatric Index of Mortality, version 3
PSG	Polysomnography
Q1	First quartile (25th percentile)
Q3	Third quartile (75th percentile)
ReLU	Rectified Linear Unit
REM	Rapid Eye Movement Sleep
RNN	Recurrent Neural Network
SWS	Slow-Wave Sleep

Introduction

Sleep is essential for maintaining both physical and mental health, and plays a vital role in enhancing quality of life and recovery from illness [1]. In critically ill paediatric patients, sleep is particularly important for both recovery and development [2–5]. However, sleep deprivation and fragmentation are common in the paediatric intensive care unit (PICU), and may impact short-term recovery and long-term neurocognitive outcomes [2]. This highlights the need to integrate sleep monitoring into clinical care, for which accurate classification of sleep is crucial to understand the physiological mechanisms of sleep [6, 7].

Polysomnography (PSG) is the gold standard for sleep assessment, recording multiple physiological signals, including electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), and electrocardiography (ECG). Among these, EEG is particularly crucial for distinguishing sleep stages, as it directly measures brain activity [8]. Sleep stages are commonly classified according to the guidelines of the American Academy of Sleep Medicine (AASM), which define five stages: Wake (W), rapid eye movement (REM) sleep, and three non-rapid eye movement (NREM) stages (N1–N3) [9]. These stages are characterised by differences in brainwave frequency, amplitude, and distinct electrophysiological features. Wakefulness is marked by high-frequency activity, while NREM sleep progresses through stages with progressively slower waves and deeper sleep. In contrast, REM sleep resembles wakefulness in brain activity but is distinguished by rapid eye movements. These stages repeat cyclically throughout the night, forming sleep cycles that typically last 90 to 120 minutes [9, 10].

Manual scoring of 30-second PSG epochs by clinical neurophysiology laborants is considered the gold standard for classification, but it is labour-intensive, subject to inter- and intra-observer variability, and time-consuming, making it impractical for real-time use in clinical settings [11–13]. However, accurate, real-time sleep stage classification could enhance clinical decision-making by minimising sleep disturbances and enabling medication adjustments that affect sleep quality [2]. To overcome the limitations of manual scoring, automated sleep scoring methods based on machine learning (ML) have been developed. These include both conventional models with manually engineered features and deep learning (DL) models, both of which have demonstrated promising classification performance, primarily in healthy adults [14].

However, the sleep patterns of neonates and children differ significantly from those of adults. Neonatal EEG is characterised by asynchrony and the absence of sleep-specific features commonly observed in adults. As a result, all NREM stages are typically grouped into a single stage, reflecting the distinct sleep architecture of neonates [9, 15, 16]. As children mature, their sleep patterns gradually evolve towards those of adults, with an increase in frequency and a more structured distribution of sleep stages, enabling easier sleep classification [17, 18]. The complexity of sleep staging increases further in critically ill patients in an intensive care unit (ICU) due to the effects of medications, underlying pathophysiology and ICU-related artefacts. For example, benzodiazepines and opioids lead to slower wave patterns, while ketamine leads to an increase in specific frequency bands [19–21]. Similarly, conditions such as encephalopathy and traumatic brain injury can result in irregular EEG activity and overall slowing of brain signals [22–25].

These challenges make automatic sleep classification particularly difficult in critically ill children. Previous studies at the PICU of Sophia Children's Hospital have investigated various conventional ML models for sleep staging on EEG data, but showed limited success. Therefore, this study aims to explore more complex, supervised ML approaches through DL feature extraction and dynamic models, to improve classification. Additionally, an unsupervised model will be used to identify underlying structures, offering deeper insights into the complex sleep dynamics of critically ill children. These objectives are structured around two key research questions, each addressed in a separate chapter. The overall pipeline of this

thesis is presented in Figure 1. The initial section describes the three datasets and the corresponding data preparations. Chapters 1 and 2 describe the feature extraction, model development, and evaluation for both supervised and unsupervised methods.

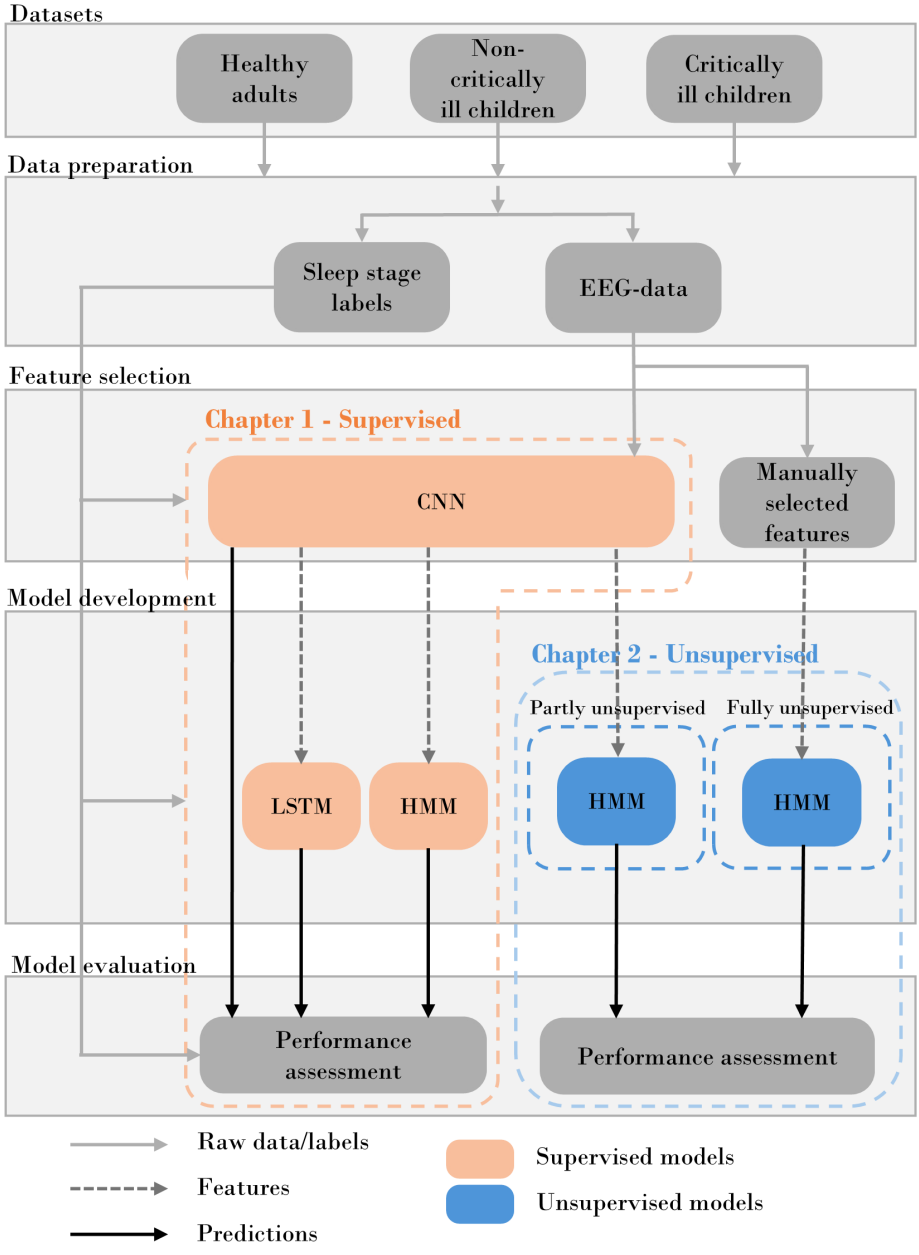


Figure 1: Overview of the thesis pipeline. The figure illustrates the flow from data preparation to model evaluation for both supervised and unsupervised approaches. EEG data and sleep stage labels from three datasets are used as input. In Chapter 1, supervised models (CNN, LSTM, and HMM) are developed. In Chapter 2, two unsupervised HMM approaches are explored: a partly unsupervised approach using CNN-derived features and a fully unsupervised approach using manually selected features. Performance is evaluated separately for each model. CNN = convolutional neural network, LSTM = long-short term memory, HMM = hidden Markov model.

Study Population and Data Acquisition

Methods

Study Population

This observational retrospective study utilised three datasets for the analysis of sleep stage classification: the Sleep-EDF, PSG, and PICU datasets.

Sleep-EDF: The Sleep-EDF dataset, an open-access resource from PhysioNet, is widely used in sleep research and enables comparisons with existing studies. It comprises PSG recordings from 20 healthy Caucasian adults aged 25–34 years [26]. One subject underwent one recording, while all others underwent two consecutive nights of recording.

PSG: This dataset includes hospital-based overnight PSG recordings from 120 non-critically ill children, collected between 2017 and 2022 at Erasmus MC Sophia Children's Hospital. Further details are described by van Twist et al. [27]. These PSG recordings were conducted for screening, diagnostic, or follow-up evaluations of conditions such as obstructive sleep apnoea. Inclusion criteria required normal physiological sleep and no atypical EEG findings. Exclusion criteria included incomplete hypnograms, poor data quality, and recordings from patients with sedative use. Participants were grouped into eight age categories: 0–2 months, 2–6 months, 6–12 months, 1–3 years, 3–5 years, 5–9 years, 9–13 years, and 13–18 years, with 15 recordings per group. For preterm children ($\leq$37 weeks gestation), age was corrected up to two years postnatal. Ethical approval with a waiver of informed consent was obtained (MEC-2021-0121).

PICU: The PICU dataset includes PSG recordings from 28 critically ill children admitted to the PICU of Erasmus MC Sophia Children's Hospital between 2020 and 2022. Recordings were conducted as part of the Critical Clock study (Netherlands Trial Register: NL8533) and the ContInNuPIC trial (Netherlands Trial Register: NL7877). The full methodologies for these studies are described by Cramer et al. and Veldscholte et al., respectively [28, 29]. Eligible participants were term-born children up to 18 years old with an expected PICU stay exceeding 48 hours. Exclusion criteria varied between studies but included recent use of melatonin or hydrocortisone and pre-existing circadian disturbances. No additional inclusion or exclusion criteria were applied for the current study. Ethical approval was obtained (MEC-2021-0121), and informed consent was acquired from participants or their guardians. Patient data, including age, gender, medical history, and medication use, were collected. Additionally, the Paediatric Logistic Organ Dysfunction Score 2 (PELOD-2) and the Paediatric Index of Mortality 3 (PIM 3) were calculated to assess illness severity.

Data Acquisition and Preprocessing

Sleep-EDF: Raw PSG signals and visually scored hypnograms were downloaded in European Data Format (EDF) from the PhysioNet database [26]. The data were sampled at 100 Hz. The Fpz-Cz EEG channel, identified in previous studies as achieving optimal performance, was selected for analysis, and no further preprocessing was applied [30]. 30-second epochs were initially scored into eight categories (Wake, N1, N2, N3, N4, REM, MOVEMENT, and UNKNOWN). However, to conform to AASM guidelines, N3 and N4 were merged into a single N3 stage. Periods of Wake at the beginning and end of recordings were shortened to include only 30 minutes before and after the sleep periods. MOVEMENT and UNKNOWN epochs were excluded, as they did not correspond to any sleep stage.

PSG and PICU: PSG recordings were conducted using BrainRT (OSG, Rumst, Belgium) or Morpheus

(Micromed Sp.A., Treviso, Italy). For the PSG dataset, recordings included eight-channel EEG, two-channel EOG, and EMG, with electrode placement following the international 10–20 system. In the PICU dataset, EEG, EOG, and EMG signals were recorded unilaterally to minimise discomfort, using the same electrode placement as in the PSG dataset. In both studies, the F3-C3 channel was utilised, as it closely resembles Fpz-Cz and has demonstrated optimal performance in a previous study [31]. For sub-analysis, EOG and EMG channels were also included.

Raw PSG signals and scored hypnograms were exported from the PSG software in EDF format. Recordings were sampled at 250 or 256 Hz but were downsampled to 100 Hz to match the Sleep-EDF data. A Butterworth band-pass filter (0.5–48 Hz) was applied to remove irrelevant frequencies [27]. Sleep stages were scored following AASM guidelines by an experienced clinical neurophysiology technician. As distinguishing between NREM sub-stages in children is challenging, epochs were classified as general NREM when specific stage classification was uncertain. Data without sleep stage labels were excluded from the analysis.

Signal analysis for all datasets was conducted in Python (3.12.7) using the following libraries: EEGlib (0.4.1.1), Hmmlearn (0.3.3), PyEDFlib (0.1.38), Scipy (1.14.1), and Tensorflow (2.18.0).

Statistics

Continuous variables are reported as means with standard deviations (SD) or medians with first and third quartiles (Q1-Q3), depending on their distribution. Categorical variables are presented as frequencies and percentages.

Results

Table 1 provides detailed information on patient characteristics, including age, sex ratio, indications for PSG, type of respiratory support, and medications administered during recordings per dataset. Details on the PSG data and the number of epochs classified into each sleep stage can be found in Table 2. The Sleep-EDF dataset consisted of 39 recordings from 20 healthy adults, with a median age of 28 (26–31) years, and 50.0% (n=10) of the participants were male. The dataset contained a total of 42,308 epochs, with a mean recording length of 9.1 (1.8) hours. The PSG dataset included 120 recordings from 120 children, with a median age of 3.5 (0.6–9.6) years, and 48.3% (n=58) were male. The mean recording length was 10.2 (1.4) hours, comprising 146,960 epochs. The PICU dataset contained 28 recordings from 28 critically ill children, with a median age of 0.3 (0.1–1.3) years, and 46.4% (n=13) were male. The mean recording length for this dataset was 22.2 (4.5) hours, resulting in 74,604 epochs.

Notably, the Sleep-EDF data contains a higher proportion of N2 sleep compared to the other datasets. In the PSG and PICU datasets, a significant portion of epochs was classified as NREM sleep, and the PICU dataset specifically showed a predominance of epochs assigned to wake and NREM stages.

Further information regarding age, diagnosis, respiratory support, neurological condition, PELOD-2 score, PIM 3 score, and medication for individual patients in the PICU dataset is available in Appendix S.1.

Table 1: Patient characteristics of the Sleep-EDF, PSG, and PICU datasets.

Patient characteristics	Sleep-EDF dataset (n = 20)	PSG dataset (n = 120)	PICU dataset (n = 28)
Age in years (median (Q1-Q3))	28 (26-31)	3.5 (0.6-9.6)	0.3 (0.1-1.3)
Sex ratio (% male (n))	50.0 (10)	48.3 (58)	46.4 (13)
PSG/PICU indication (% of patients (n))	n/a	Airway obstruction: 48.3 (58) Central sleep apnoea: 13.3 (16) Neuromuscular disease: 29.2 (35) Pulmonary disease: 9.2 (11)	Abdominal surgery: 3.6 (1) Cardiac (surgery): 39.3 (11) Infectious: 10.7 (3) Metabolic: 3.6 (1) Neurological: 10.7 (3) Oncological: 3.6 (1) Respiratory: 25.0 (7)
Respiratory support during PSG (% of patients (n))	None	None	None: 10.7 (3) Nasal cannula: 17.9 (5) non-invasive ventilation: 3.6 (1) invasive ventilation: 67.9 (19)
Sedative/analgesic medication during PSG (% of patients (n))	None	None	Esketamine: 25.0 (7) Midazolam: 64.3 (18) Opioids: 67.9 (19)

Q1 = first quartile, Q3 = third quartile, PSG = polysomnography, PICU = paediatric intensive care unit.

Table 2: Polysomnography data characteristics of the Sleep-EDF, PSG and PICU datasets.

PSG data characteristics	Sleep-EDF dataset (n = 39)	PSG dataset (n = 120)	PICU dataset (n = 28)
Length of PSG recording in hours (mean (SD))	9.1 (1.8)	10.2 (1.4)	22.2 (4.5)
Total amount of epochs (n)	42,308	146,960	74,604
Amount of epochs for each stage (% of epochs (n))			
Wake	19.5 (8,285)	19.9 (29,373)	34.2 (25,519)
REM	18.2 (7,717)	8.1 (11,830)	9.5 (7,070)
N1	6.6 (2,804)	7.7 (11,311)	4.4 (3,265)
N2	42.0 (17,799)	20.0 (29,432)	9.2 (6,870)
N3	13.5 (5,703)	24.2 (35,578)	6.2 (4,647)
NREM	None	20.0 (29,436)	36.5 (27,233)

PSG = polysomnography, SD = standard deviation, REM = rapid eye movement, NREM = non rapid eye movement, N1-N3 = non rapid eye movement stage 1-3.

Chapter 1: Complex Machine Learning Approach

1.1 Research Question

Can deep learning-based feature extraction combined with dynamic machine learning models effectively address the challenges of automatic sleep staging using EEG data in non-critically ill and critically ill children?

As mentioned earlier, ML models, including simpler approaches, have demonstrated strong performance in automatic sleep staging for healthy adults. For the Sleep-EDF dataset, many studies reported accuracies exceeding 80% for classification of the five sleep stages [14]. Figure 2 illustrates why this is feasible: distinct differences between sleep stages are visually evident in EEG data. These differences are characterised by variations in frequency, amplitude, and specific features. Wakefulness is characterised by low-amplitude, high-frequency alpha waves (8–13 Hz) and beta waves (>13 Hz). N1 sleep is marked by the transition from alpha to theta waves (4–7 Hz) and low voltage. N2 sleep is dominated by theta waves, with the defining presence of sleep spindles and K-complexes, as marked in Figure 2. N3 sleep, representing deep sleep, is distinguishable by high-amplitude, low-frequency delta waves (0.5–2 Hz). REM sleep exhibits low-amplitude mixed-frequency activity, resembling wakefulness but with bursts of rapid eye movements, reflecting intense brain activity and dreaming [9]. Typically, sleep follows a characteristic cyclical pattern, progressing through these stages before repeating the cycle. A normal cycle generally moves from light NREM sleep to deep NREM sleep, then transitions back through lighter NREM stages before entering REM sleep, after which the cycle begins again [9, 10].

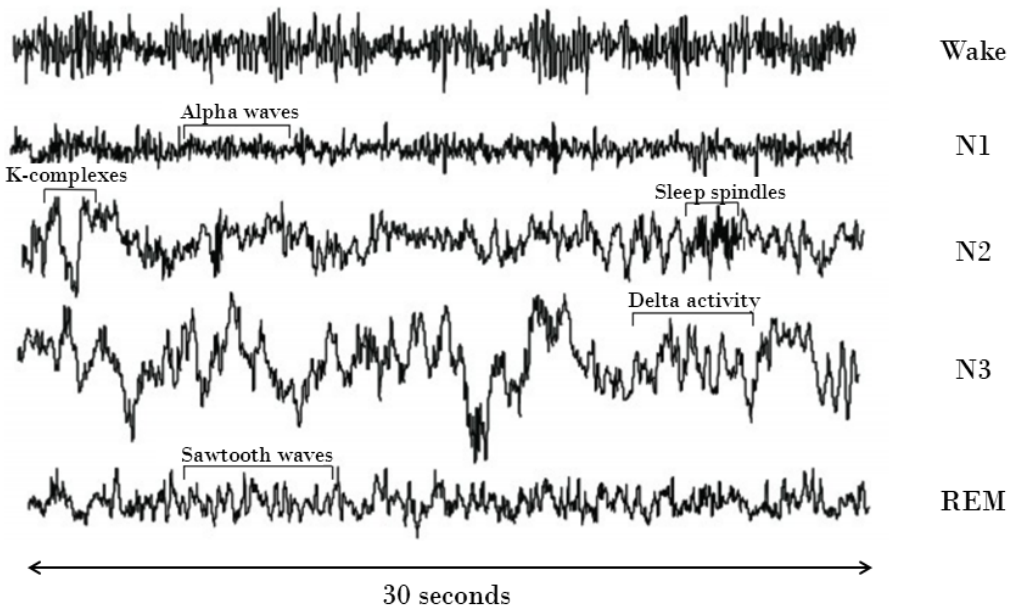


Figure 2: Example of electroencephalography signals for all sleep stages in a healthy adult. Visually distinct characteristics across different sleep stages are shown.

In contrast, significantly fewer studies have investigated automatic sleep staging in children and critically ill patients. At Sophia Children's Hospital, a previous thesis employed EEG-based index measures and

conventional machine learning models to classify sleep stages in both non-critically and critically ill children. In non-critically ill children, an XGBoost model achieved reasonably good performance, with a maximum accuracy of 79% when classifying three stages. In contrast, performance in critically ill children was limited, with the same model yielding a maximum accuracy of only 55% [31]. These results indicate that sleep stage classification in children presents greater challenges than in adults, with classification in critically ill children proven to be particularly difficult. To illustrate the unique challenges posed by this population, Figure 3 presents an example of EEG data from a critically ill child. Unlike the clear distinctions observed in a healthy adult (Figure 2), the differences between Wake, NREM, and REM stages are far less pronounced. As also shown in the figure, all three NREM stages are often merged into a single category due to the difficulty in distinguishing them, particularly in young children or in the presence of abnormal sleep patterns.

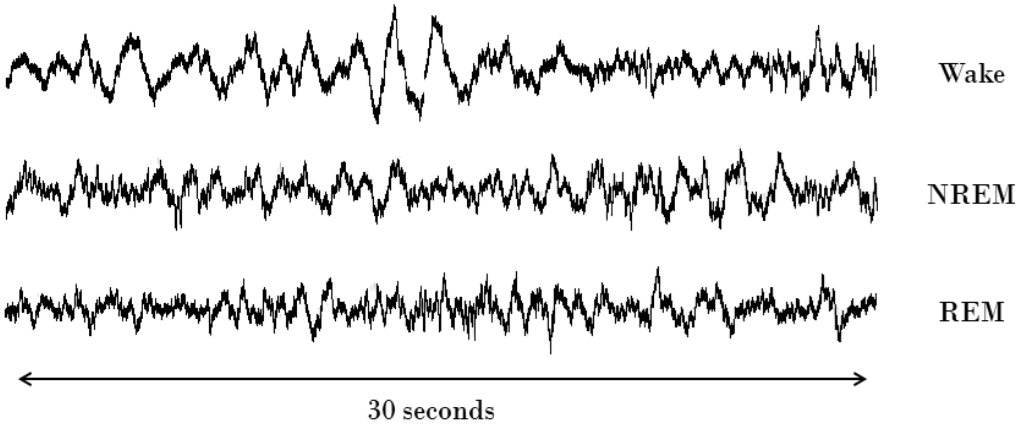


Figure 3: Example of electroencephalography signals for sleep stages in a critically ill child. Visually less distinguishable differences across sleep stages are shown.

Traditional ML models with manually selected features can struggle to capture the complex patterns present in EEG data, particularly when sleep stages are less distinct [32–34]. To address this limitation, this thesis aims to investigate whether a DL model can improve sleep stage classification by automatically extracting features directly from EEG data. Specifically, a convolutional neural network (CNN) will be employed for feature extraction, as CNNs have proven effective in recognising complex patterns [35, 36]. The hypothesis is that the CNN will better capture subtle differences in EEG signals, leading to improved performance in sleep staging for non-critically ill and critically ill children.

After extracting features through the CNN, this study will further explore whether incorporating the sequential nature of sleep stages enhances classification performance. Dynamic machine learning models, such as recurrent neural networks (RNNs) and hidden Markov models (HMMs), are well-suited for modelling the temporal structure of sequential data. Previous studies have demonstrated their high performance in sleep staging using adult datasets [35, 37–39]. Long short-term memory (LSTM) networks, a type of RNN, are commonly used due to their ability to capture both short- and long-term temporal dependencies. HMMs, although simpler and designed to model only short-term dependencies, have demonstrated comparable performance, making them a viable alternative [14]. Therefore, both an LSTM network and HMM will be developed and evaluated in this thesis. The hypothesis is that integrating spatial features extracted by the CNN with temporal dependencies captured by the LSTM or HMM will result in enhanced performance compared to traditional approaches for sleep staging in non-critically ill and critically ill children.

1.2 Methods

1.2.1 Model Development

Convolutional Neural Network

For this study, a CNN based on the design proposed by Supratak et al. was constructed, of which the structure is shown in Figure 4 [30]. The CNN was trained to learn filters capable of extracting time-invariant features from individual 30-second, single-channel EEG epochs. In Supratak et al., two separate CNN pathways were employed, one with small filters and the other with large filters in the first convolutional layer. This design was inspired by signal processing techniques, which balance the trade-off between spatial and frequency resolution during feature extraction. The small filter pathway is more effective in capturing spatial information (i.e. when certain patterns appear), whereas the large filter pathway is better suited to extracting frequency-based features. Both pathways consist of four convolutional layers and two max-pooling layers. Each convolutional layer sequentially performs three operations: convolution with its filters, batch normalisation, and activation via the rectified linear unit (ReLU) function. The pooling layers downsample the feature maps, reducing spatial dimensions while preserving essential features.

The filter sizes, number of filters, stride sizes, and pooling sizes match the configurations outlined by Supratak et al. and are illustrated in Figure 4. After processing, the outputs of the two CNN pathways are concatenated to create a combined representation of the EEG epoch. These combined feature vectors are either passed through a dense layer with a softmax activation function to generate the output probabilities of the CNN alone, or used directly as input features for the LSTM or HMM. A detailed explanation of CNN architecture and its components is provided in Appendix S.2.

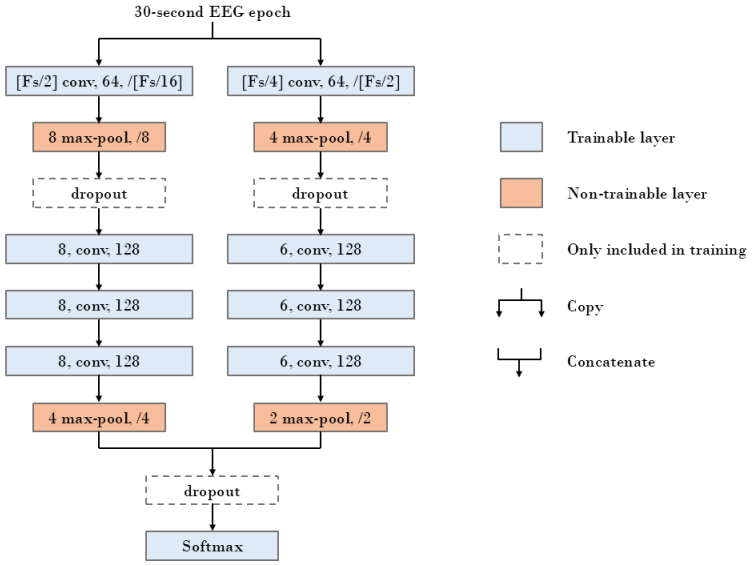


Figure 4: Architecture of the convolutional neural network. The model processes 30-second EEG epochs using two parallel convolutional branches. Outputs of both branches are concatenated and passed through a final layer for the prediction of sleep stages. Each convolutional block shows a filter size, the number of filters, and a stride size. Each max-pooling block shows a pooling size and stride size.

Long Short-Term Memory

As the first method for modelling sequential dependencies, an LSTM network was employed. The architecture used in this study is also adapted from the design proposed by Supratak et al. and is illustrated in Figure 5 [30]. While the original design incorporated a bidirectional LSTM (Bi-LSTM), a review of comparative studies indicated that Bi-LSTMs generally do not significantly outperform standard LSTMs in sleep stage classification tasks, yet they approximately double the computational time [14]. Consequently, the Bi-LSTM was replaced with a standard LSTM. Utilising the CNN outputs for a sequence of epochs, the model incorporates two LSTM layers to capture temporal dependencies. This enables the model to use information from previous sleep epochs to determine the current sleep stage. In parallel, a fully connected layer processes the features extracted by the CNN into a vector, which is element-wise added to the output of the LSTM layers. Finally, the combined output, containing both spatial and temporal features, is passed through a dense layer with a softmax function to generate output probabilities. A detailed explanation of the model architecture and its parameters is provided in Appendix S.3.

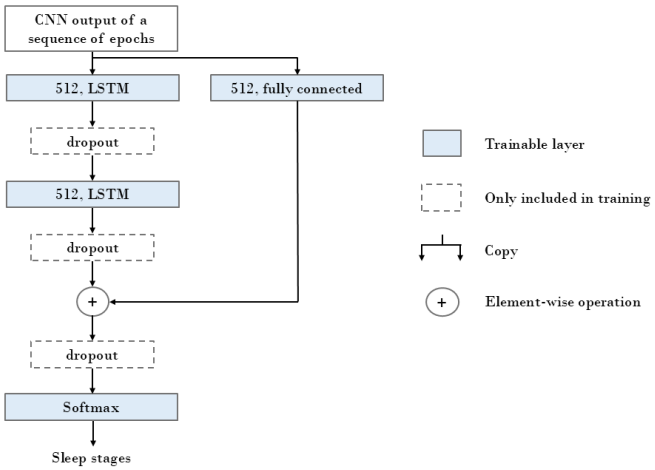


Figure 5: Architecture of the long short-term memory (LSTM) model. It utilises the output for a sequence of epochs of the convolutional neural network as input to two LSTM layers to capture sequential dependencies. By performing an element-wise operation with the output of the fully connected layer, the temporal and spatial features are combined for the classification of the sleep stages. The LSTM and fully connected layers both show the hidden size.

Hidden Markov Model

As the second dynamic model, an HMM was designed to model the temporal dynamics of sleep stage transitions through stochastic processes. A simplified version of this architecture is illustrated in Figure 6, which shows how the model uses transition probabilities to represent the likelihood of moving between hidden states (black arrows) and emission probabilities to model the relationship between each hidden state and the observed features (grey arrows). The hidden states correspond to different sleep stages, while the observed features are features derived from the CNN. Two feature extraction methods were employed: the first used the probability outputs from the CNN's softmax layer (HMM-s), while the second applied normalisation and Principal Component Analysis (PCA) to the features from the final layer before softmax activation (HMM-f). In the HMM-s, the emission probabilities were modelled as discrete distributions, whereas in the HMM-f, they were modelled as Gaussian distributions. The distribution of the first principal components per sleep stage for each dataset are visualised in Appendix S.4 to assess the suitability of Gaussian emission assumptions. A detailed explanation of the model is provided in Appendix S.5.

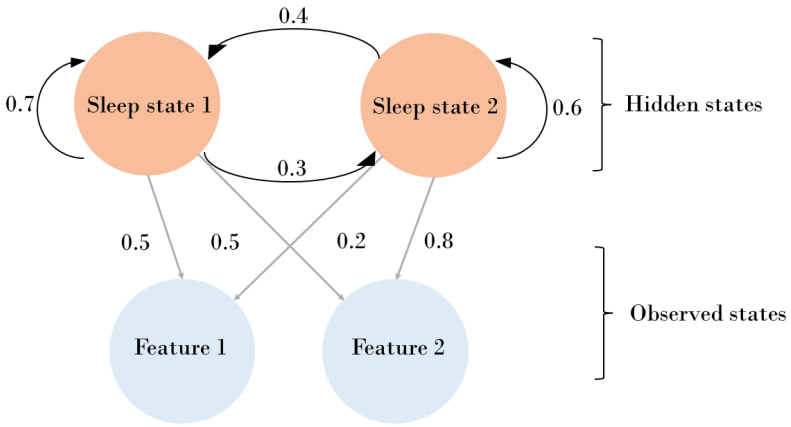


Figure 6: Architecture of the Hidden Markov Model. It consists of hidden states representing different sleep stages and observed states representing features derived from the CNN. Transition probabilities between hidden states (black arrows) and emission probabilities from hidden to observed states (grey arrows) are modelled.

1.2.2 Model Training

Convolutional Neural Network

A supervised training algorithm was employed to train the CNN using a class-balanced training dataset to limit overfitting on the majority sleep stages. The class-balanced dataset was created by duplicating epochs of minority sleep stages to match the number of epochs in the most prevalent sleep stage. The model's trainable parameters (i.e. layer weights and biases) were updated using the Adam optimiser, a gradient-based algorithm that dynamically adapts the learning rate for each parameter. Based on Supratak et al., an initial learning rate of 1×10^{-4} was chosen [30]. The categorical cross-entropy loss function was used to penalise incorrect predictions proportionally to their confidence, guiding the model toward more accurate classifications. Training was performed in mini-batches of 100 sleep epochs, where each mini-batch was used to compute the loss and update the trainable parameters accordingly. The training process was governed by several hyperparameters, including a maximum of 100 epochs and the use of early stopping: training was halted if validation accuracy did not improve for 20 consecutive epochs. The trainable parameters corresponding to the highest validation accuracy were saved for subsequent evaluation.

Long Short-Term Memory

The LSTM was also trained using a supervised algorithm; however, the original class-imbalanced training dataset was used. Oversampling to balance the dataset was not feasible, as the temporal sequence of epochs is crucial for the LSTM. The input data for the LSTM was first passed through the saved CNN, thus with the best trainable parameters. The LSTM was subsequently trained using the Adam optimiser to update the trainable parameters. An initial learning rate of 1×10^{-5} was chosen, again based on Supratak et al [30]. Although experiments were conducted using a class-imbalanced loss function, this did not result in performance improvements; therefore, the standard cross-entropy loss function was used for optimisation. Similar to the hyperparameters utilised for the CNN, a maximum of 100 training iterations was used and early stopping was applied. Trainable parameters were saved based on the highest validation accuracy. The LSTM was trained with varying sequence lengths of input epochs to investigate this effect on performance. The batch size was adjusted inversely to the sequence length to ensure computational efficiency.

Hidden Markov Model

As with the LSTM, oversampling to balance the dataset was not feasible, and the original training dataset was utilised. The transition and emission probabilities were estimated from the dataset and remained fixed throughout the process, eliminating the need for the Baum-Welch algorithm. The Viterbi algorithm was applied to determine the most likely sequence of hidden states during both training and testing, ensuring an optimal sleep stage sequence given the observed data. A detailed explanation of these algorithms is provided in Appendix S.5.

1.2.3 Regularisation

To address overfitting in both the CNN and LSTM, two regularisation techniques, as outlined by Supratak et al., were employed [30]. The first regularisation method is dropout, which was solely applied during training. Dropout randomly sets a fraction of the units in the previous layer to zero, thereby removing them along with their connections. This prevents the model from relying too heavily on specific features, encouraging it to learn more robust and generalised patterns. The dropout layers are depicted in Figures 4 and 5.

The second regularisation method is L2 weight decay, which was applied to the first layer of both pathways in the CNN. L2 weight decay introduces a penalty term to the loss function, discouraging the model from assigning excessively large weights. Following Supratak et al., L2 weight decay was specifically applied to the first convolutional layers to prevent the model from overfitting to noise and artefacts [30]. It was not applied to subsequent layers to avoid overly constraining the model to learn.

1.2.4 Model Development

All models were evaluated using k-fold cross-validation (CV), where k was set to 20 for the Sleep-EDF dataset and 5 for both the PSG and PICU datasets. Specifically, in each fold, recordings from $N_s - \frac{N_s}{k}$ subjects were used to train the model, and the remaining $\frac{N_s}{k}$ subjects were used to test the trained model, where N_s is the total number of subjects in the dataset. This process was repeated k times, ensuring that all recordings were included in the testing set once. After completing the CV procedure for all folds, the performance metrics per fold are combined for the overall performance.

1.2.5 Classification Tasks and Performance Metrics

Performance metrics for all models were determined across the three datasets, with an additional category comprising data from children aged 9 to 18 years from the PSG dataset (30 patients). This category was included due to the more mature sleep patterns observed in this age group, allowing for a better comparison with the Sleep-EDF dataset. Models were developed for both three-state and five-state sleep stage classification. The three-state classification consists of Wake, REM, and a general NREM category, and was applied to all four datasets. The five-state classification, including Wake, REM, N1, N2, and N3 stages, was applied to the Sleep-EDF dataset, to 92 patients of the PSG dataset (as the remaining 28 recordings did not contain separate NREM stages), and to children aged 9 to 18 years from the PSG dataset.

The models were evaluated using several performance metrics, including macro-averaged F1 scores (MF1), overall accuracy (ACC), Cohen's Kappa (kappa), and the area under the receiver operating characteristic curve (AUC ROC). Both MF1 and kappa scores account for class imbalance, with MF1 providing a balanced measure of precision and recall across all classes, and κ measuring agreement while considering

the possibility of chance-level agreement. For each performance metrics the mean and standard deviation over all folds were provided. Detailed calculations of these metrics are provided in Appendix S.6.

1.3 Results

1.3.1 Parameter Tuning and Input Data Adaption

(Hyper)parameter Tuning

As mentioned, this study employed a CNN and LSTM architecture similar to that used by Supratak et al. [30]. The training hyperparameters, such as the learning rate used by the Adam optimiser, were set according to the values reported to yield the best validation performance in their study. No additional tuning of hyperparameters was conducted. The model's trainable parameters (i.e. the weights and biases within the convolutional and LSTM layers) were learned from the data during training.

Adapting the Input Channels

As previously mentioned, for the PSG and PICU datasets, the F3-C3 EEG channel was selected as it closely resembles the Fpz-Cz channel used in the study by Supratak et al. and has demonstrated the strongest one-channel performance in a prior study [30, 31]. However, in clinical practice, important information for sleep staging is also obtained from EOG and EMG channels, particularly for accurately identifying REM sleep [9]. Therefore, an evaluation was conducted on the PSG dataset to determine whether the inclusion of these channels improved CNN performance. However, no considerable performance enhancement was observed and further evaluations solely utilised the F3-C3 EEG channel. The results of this evaluation are provided in Appendix S.7.

Adapting the Sequence length

The effect of sequence length (amount of epochs) as input of the LSTM on its performance was assessed for three-state classification using the PSG dataset, with results presented in Appendix S.8. Performance metrics remained consistent across sequence lengths of 5, 10, 20, 60, 120, and 240 epochs, suggesting that sequence length does not have a major impact on performance. As Supratak et al. used a sequence length of 20 epochs, this was adopted for further evaluations in all datasets [30].

Adapting the Number of Principal Components

The impact of the number of principal components utilised as input for the HMM was assessed using the Sleep-EDF dataset, with results presented in Appendix S.9. The explained variance of the principal components for all datasets is provided in Appendix S.10. An analysis of the explained variances revealed that to achieve a high variance, a relatively large number of principal components is necessary. However, performance remained largely stable when selecting between 10 and 500 components. To balance the trade-off between the explained variance and reducing computational time, which increases with more components in the HMM, a total of 50 principal components was selected for further evaluation.

1.3.2 Final Results

Table 3 summarises the performance metrics for three-state and five-state sleep stage classification across the Sleep-EDF, PSG, and PICU datasets.

Table 3: Performance metrics for three- and five-state sleep stage classification for the Sleep-EDF, PSG and PICU dataset. Results were obtained by 20- and 5-fold cross-validation for the Sleep-EDF and PSG/PICU dataset, respectively. The results in italic correspond to the patients from the PSG dataset for whom five sleep stages were labelled.

		Three-state performance in % (SD)				Five-state performance in % (SD)			
Dataset	**Model**	**Acc**	**MF1**	**kappa**	**AUC**	**Acc**	**MF1**	**kappa**	**AUC**
Sleep-EDF	CNN	88.3 (5.7)	85.4 (6.8)	78.0 (9.6)	96.8 (3.0)	80.2 (7.4)	74.0 (7.0)	72.8 (9.5)	95.1 (2.3)
(n = 39)	LSTM	88.3 (6.1)	85.2 (7.6)	77.9 (10.4)	96.8 (3.1)	81.6 (7.0)	74.0 (7.7)	74.5 (9.3)	95.1 (2.3)
	HMM-s	90.2 (5.9)	87.8 (7.2)	81.7 (10.1)	95.2 (4.4)	83.5 (6.7)	76.5 (6.3)	77.1 (8.7)	93.2 (3.5)
	HMM-f	86.4 (4.8)	83.7 (5.7)	75.6 (8.3)	94.9 (7.1)	80.1 (7.1)	73.2 (8.2)	72.7 (9.3)	94.5 (2.7)
PSG - 9 to	CNN	80.6 (4.3)	75.6 (4.6)	65.7 (6.1)	93.3 (2.4)	67.4 (4.5)	62.4 (4.5)	58.6 (5.3)	90.7 (2.4)
18 years	LSTM	78.4 (4.8)	71.6 (6.5)	59.6 (8.5)	91.6 (3.5)	64.9 (5.3)	55.8 (6.5)	54.6 (7.0)	88.5 (7.0)
(n = 30)	HMM-s	79.2 (7.2)	73.8 (8.1)	63.2 (10.8)	89.4 (4.8)	64.7 (7.0)	58.7 (6.7)	55.1 (8.5)	86.9 (3.8)
	HMM-f	72.4 (5.7)	66.4 (5.6)	53.7 (7.1)	88.7 (3.2)	59.3 (7.1)	55.3 (6.5)	48.7 (7.9)	86.9 (2.8)
PSG	CNN	71.3 (1.7)	69.2 (1.9)	54.2 (2.4)	89.4 (1.0)	*62.2 (3.9)*	*57.4 (2.4)*	*52.0 (4.5)*	*87.6 (1.2)*
(n = 120)	LSTM	77.4 (0.7)	73.5 (1.0)	61.4 (1.1)	91.1 (0.7)	*66.0 (2.1)*	*56.5 (2.1)*	*56.1 (2.6)*	*88.8 (0.9)*
(n = 92)	HMM-s	73.1 (1.5)	68.8 (2.2)	54.7 (2.8)	85.6 (2.8)	*64.0 (2.4)*	*57.8 (2.9)*	*53.8 (2.9)*	*85.2 (1.5)*
	HMM-f	67.7 (2.1)	64.9 (1.7)	49.0 (2.6)	86.3 (1.0)	*61.7 (1.2)*	*56.6 (1.7)*	*51.2 (1.4)*	*86.5 (0.7)*
PICU	CNN	59.1 (5.0)	45.9 (9.1)	26.5 (10.9)	69.3 (10.3)	-	-	-	-
(n = 28)	LSTM	61.4 (6.2)	43.2 (6.4)	26.1 (12.8)	67.1 (6.2)	-	-	-	-
	HMM-s	52.7 (6.1)	39.1 (6.4)	17.5 (6.6)	62.4 (6.6)	-	-	-	-
	HMM-f	44.6 (15.2)	35.2 (8.8)	9.1 (9.7)	59.2 (8.9)	-	-	-	-

SD = standard deviation, Acc = accuracy, MF1 = macro-F1 score, kappa = Cohen's kappa, AUC = area under the receiver-operating characteristic curve, CNN = convolutional neural network, LSTM = long short-term memory, HMM-s = hidden Markov model-softmax layer, HMM-f = hidden Markov model-final layer, - : model was not trained on the dataset.

Overall, all models performed strongly on the Sleep-EDF dataset, with mean accuracies ranging from 86.4% to 90.2% for three-state classification and from 80.1% to 83.5% for five-state classification. The corresponding mean MF1 scores were approximately 85% and 75% for three- and five-state classification, respectively, reflecting a good performance with balanced class distributions.

Compared to the Sleep-EDF dataset, a decline in performance is observed in the PSG subgroup of patients aged 9 to 18 years. For three-class sleep stage classification, the maximum achieved accuracy and MF1 score are 80.6% and 75.6%, respectively. This decline is more pronounced in the five-class classification, with a maximum accuracy of 67.4% and an MF1 score of 62.4%.

Sleep staging appears to be even more challenging in the full PSG dataset. Mean accuracies range from 67.7% to 77.4% for three-state classification and from 61.7% to 66.0% for five-state classification. Although AUC-ROC values remain relatively high (i.e. close to 90%), the decline in performance is also evident in the MF1 and kappa score.

Performance declines further when critically ill children from the PICU dataset are considered. The models achieve mean accuracies ranging from 44.6% to 61.4%, with low MF1 scores between 35.2% and 45.9%, and kappa values ranging from 9.1% to 26.5%. ROC-AUC scores are also substantially lower than for the other datasets, ranging between 59.2% and 69.3%.

The LSTM and HMM models do not demonstrate a consistent overall improvement, but rather show

varied effects across different metrics and datasets. In the Sleep-EDF dataset, incorporating temporal information yields only minor differences in performance across metrics. In contrast, in the full PSG dataset, the inclusion of sequential dependencies using an LSTM improves performance, with the mean accuracy and MF1 score increasing from 71.3% to 77.4% and from 69.2% to 71.6%, respectively, for three-class classification. The HMM-s model, which uses the softmax output probabilities, also improves accuracy to 73.1%, although it leads to a drop in the MF1 score to 68.8%. The HMM-f model, which uses features from the final pre-softmax layer, consistently performs worse than the other methods across all datasets. In the PICU dataset, the inclusion of temporal information appears to further degrade performance, with both HMM models performing particularly poorly.

Within the Sleep-EDF, somewhat higher SDs are observed, which can be attributed to the use of 20-fold CV, equivalent to leave-one-subject-out CV. This approach increases the likelihood of greater SDs due to variability between subjects. In contrast, the whole PSG dataset exhibits lower SDs, indicating consistent performance across folds and minimal variability between them. The higher SDs observed in the PICU dataset suggest greater variability across different CV-folds.

1.4 Analysis of the Results

This section provides a detailed examination of model performance across the different datasets, and further analyses are conducted to enhance understanding of the underlying data characteristics.

1.4.1 Sleep Stage-Specific Classification Performance

Table 4 presents the confusion matrix of the CNN trained on the PSG dataset, including the F1 score for each sleep stage. The results indicate that N1 is the most difficult stage to classify, which is consistent with findings in previous studies [14, 30]. This challenge arises because N1 is a transitional stage between wakefulness and N2, meaning it shares characteristics of both sleep and wake stages [9]. Consequently, a considerable number of wake, N2, and REM epochs are misclassified as N1. Additionally, a substantial proportion of wake epochs are misclassified as REM sleep, likely due to similar low-amplitude and mixed-frequency EEG patterns. Thereby, REM sleep is characterised by eye movements and occasional muscle activity, which can introduce artefacts resembling the high-frequency activity typically observed during wake [40]. N2, as an intermediate sleep stage between N1 and N3, is frequently misclassified as either N1 or N3. Interestingly, a notable proportion of N2 epochs are also misclassified as REM sleep, the cause of which remains unclear. Lastly, N3 is classified with the highest F1 score, likely due to the distinctive pattern of deep sleep, which features high-amplitude, low-frequency delta waves, making it distinguishable from other sleep stages [9]. Confusion matrices for all datasets are provided in Appendix S.11.

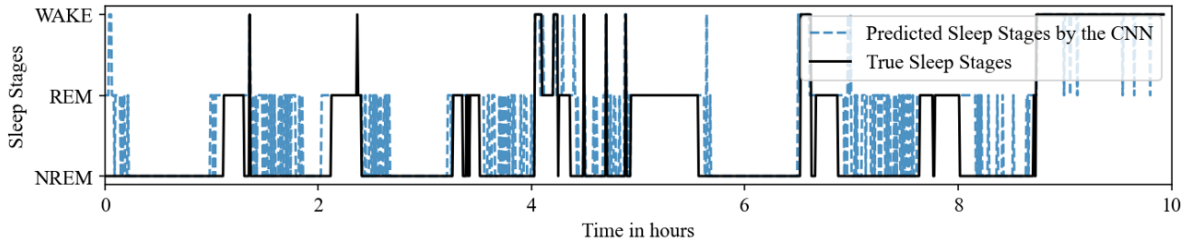
Table 4: Confusion matrix of the convolutional neural network on the PSG dataset for five-state classification. Results were obtained using 5-fold cross-validation. The amount of labels represent the total across all folds. F1-scores for each sleep stage are reported as the mean (SD) across folds.

Actual Labels	Predicted Labels					Mean F1 in % (SD)
	Wake	N1	N2	N3	REM	
Wake	13258	2829	380	225	3518	71.1 (4.8)
N1	1544	2664	858	165	5535	21.1 (4.5)
N2	866	4091	13205	3881	5224	55.7 (7.3)
N3	240	1008	4550	23198	2259	78.4 (2.8)
REM	1022	2860	312	341	16167	60.6 (5.2)

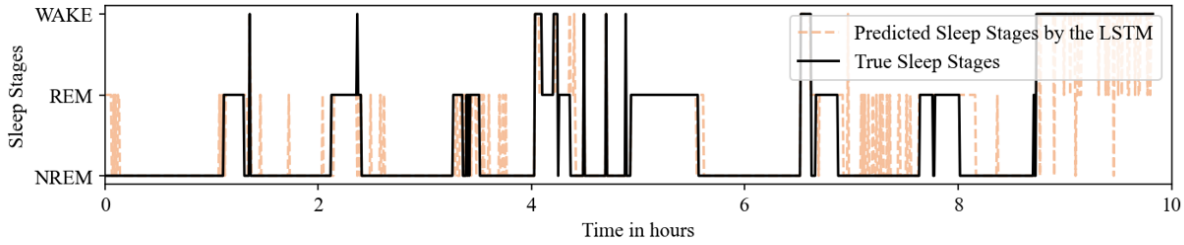
SD = standard deviation, N1-N3 = non rapid eye movement stages 1-3, REM = rapid eye movement.

1.4.2 Comparison of Model Predictions Using Hypnograms

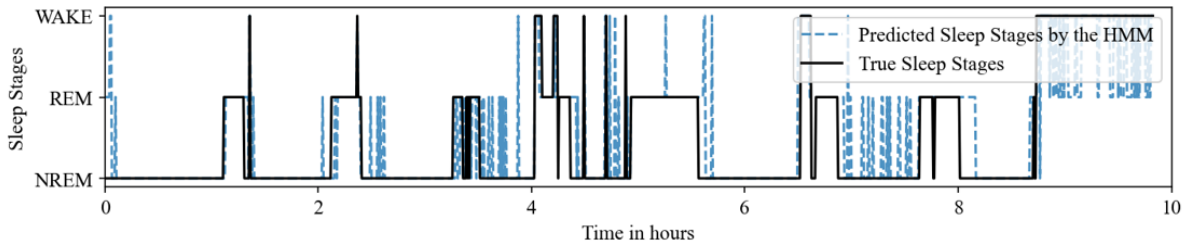
Figure 7 presents hypnograms illustrating the agreement between manually scored sleep stages and predictions made by the CNN, LSTM, and HMM-f for a patient from the PSG dataset. Overall, visually a good level of agreement is achieved; however, short-term inconsistencies can be observed in the classification of Wake and NREM states. When comparing the LSTM and HMM-f to the CNN, a noticeable smoothing effect is observed. This suggests that incorporating temporal dependencies into the model can lead to a more stable classification over time, reducing abrupt transitions between sleep stages.



(a) Convolutional neural network. The accuracy is 76.4% and the macro-F1 is 78.7%.



(b) Long short-term memory. The accuracy is 87.2% and the macro-F1 is 84.8%.



(c) Hidden Markov model-final layer. The accuracy is 83.3% and the macro-F1 is 78.3%.

Figure 7: Comparison of manually scored and model-predicted hypnograms for three-state classification in a patient from the PSG dataset. REM = rapid eye movement, NREM = non rapid eye movement, CNN = convolutional neural network, LSTM = long short-term memory, HMM = hidden Markov model.

1.4.3 Analysis of Performance on the PSG Dataset

Comparison to Performance on the Sleep-EDF Dataset

As the brains of children aged 9 to 18 years are more matured, their sleep patterns are expected to be comparable to those observed in the adult Sleep-EDF dataset [9]. However, when comparing the two datasets in Table 3, a substantial difference in performance is observed. Although the children in the PSG dataset are not critically ill, many present with airway obstructions, sleep apnoea, or neuromuscular conditions, which may influence sleep architecture. This likely results in a more heterogeneous dataset compared to the Sleep-EDF dataset, thereby increasing the complexity of accurate sleep stage classification. For instance, a study reported an 8% decrease in classification accuracy between patients without obstructive sleep apnoea syndrome (OSAS) and those with severe OSAS [41]. Notably, the lowest MF1

score in this subgroup for an individual patient was 27.3%, which was for a patient with Crouzon syndrome, a genetic disorder characterised by premature fusion of certain skull bones. This implies that such patient characteristics make accurate classification more challenging.

Influence of Age on Classification Performance

In Table 5, the left side presents the mean (SD) scores for the eight different age groups when the CNN is trained on the entire PSG dataset. The highest performance is observed in the middle age groups (6 months to 9 years), possibly because EEG patterns in the youngest and oldest children can be seen as outliers of the whole dataset, due to either a very immature or mature brain, making them more challenging for the model to classify accurately. In contrast, the middle-aged children exhibit sleep patterns that fall between those of younger and older children, therefore aligning more closely with the overall characteristics in the dataset. As expected, the lowest performance is observed in the youngest age groups, likely due to their less distinct sleep stages [9, 15].

On the right side of the Table 5, the results are shown for when the CNN is trained separately for four groups, each containing 30 patients. This results in improved performance for both the youngest and oldest age groups. This is likely because all patients in these subsets either have a very immature or matured brain, allowing the model to learn more representative features for each subset. This improvement is particularly pronounced in the oldest age category, likely due to the more mature and distinct sleep states. For the subgroups of patients aged between 3 and 9 years the performance difference is not substantial, with a decrease in accuracy and an increase in the MF1 score. However, a decrease in performance is noted for patients aged between 6 months and 3 years, possibly due to the reduced amount of training data compared to training on the whole dataset, leading to insufficient generalisation across variations within these age ranges. Thereby, a high standard deviation is observed, further suggesting substantial heterogeneity within this subgroup. This aligns with the rapid developmental changes in sleep architecture occurring during this period, which lead to greater variability in sleep stage characteristics [17, 31].

Table 5: Performance of the convolutional neural network across different age groups of the PSG dataset. The results were obtained by 5-fold cross-validation for three-state classification. The left side presents results per age category after training on the whole dataset. The right side presents results per two age categories when trained on that subset of the dataset. All age categories consist of 15 patients.

	Mean (SD) three-state performance in %			
	Training on the whole dataset		Training per age category	
Age category	**Acc**	**MF1**	**Acc**	**MF1**
0-2 Months	60.7 (11.3)	59.3 (12.0)	70.1 (2.9)	68.8 (3.2)
2-6 Months	67.0 (17.1)	63.2 (17.7)		
6-12 Months	75.3 (9.0)	73.0 (10.3)	60.2 (11.4)	55.4 (12.7)
1-3 Years	77.4 (8.4)	73.2 (8.7)		
3-5 Years	78.3 (13.3)	68.4 (16.1)	73.5 (6.2)	69.9 (7.2)
5-9 Years	77.6 (9.4)	68.2 (10.3)		
9-13 Years	70.9 (18.7)	63.1 (17.5)	80.6 (4.3)	75.6 (4.6)
13-18 Years	70.6 (9.3)	63.8 (10.2)		

SD = standard deviation, Acc = accuracy, MF1 = macro-F1 score.

1.4.4 Analysis of Performance on the PICU Dataset

Influence of Patient Characteristics on the Performance

Appendices S.12 and S.13 provide detailed insights into individual patient characteristics and per-patient performance for the PICU dataset. Appendix S.12 visualises the percentage of epochs classified into each sleep stage for individual patients. This reveals substantial variability between patients: while some exhibit comparable proportions of Wake, NREM, and REM stages, others have nearly all epochs classified as NREM sleep. Such imbalances may complicate model training. However, the influence is difficult to establish, while no consistent relationship between these distributions and performance metrics was observed.

Appendix S.13 provides further insight into individual patient characteristics, allowing for an examination of their potential influence on model performance. However, among patients diagnosed with encephalopathy, macro-F1 scores range from 24.8% to 67.4%, making it difficult to identify a clear impact of this condition on classification performance. Similarly, the two patients with the highest PELOD-2 scores (14 and 21) yielded macro-F1 scores of 61.8% and 32.6%, respectively, again showing no consistent pattern. With regard to age, the six oldest patients (aged 5 to 18 years) all demonstrated low performance, with macro-F1 scores ranging from 22.9% to 38.1%. Although this negative association is unexpected, it may be attributed to the young median age of the dataset (95.0 days). Overall, the high variability in clinical and sleep characteristics across patients in this small and heterogeneous dataset makes it challenging to draw definitive conclusions about the influence of individual factors. Moreover, a 'typical' PICU patient is difficult to define, as each patient profile differs considerably from the next, making the evaluation of the effect of individual characteristics challenging.

Inclusion of EOG and EMG Channels

The addition of EOG and EMG channels as input did not improve performance for the PSG dataset. However, in clinical practice, these signals play a crucial role in sleep stage classification, particularly is cases where the stages are not well-defined. Therefore, the inclusion of these channels as input for the CNN was evaluated for the PICU dataset to determine whether they could enhance performance. The results, presented in the first two rows of Table 6, indicate a modest improvement in performance. However, a paired t-test yields a p-value of 0.3702, suggesting that the improvement is not consistent across all patients. Moreover, with a mean macro-F1 score of 50.1% and a mean kappa of 33.7%, overall performance remains limited.

Table 6: Performance metrics for three-state sleep stage classification for the PICU dataset with varying input channels and dataset size. Results were obtained by 5-fold cross-validation. The subset (last row) consists solely of patients achieving a MF1 of >35.0% when training on the whole dataset.

	Mean three-state performance in % (SD)			
Input Data	**Acc**	**MF1**	**Kappa**	**AUC**
EEG (n = 28)	59.1 (5.0)	45.9 (9.1)	26.5 (10.9)	69.3 (10.3)
EEG, EOG, and EMG (n = 28)	63.2 (3.2)	50.1 (5.2)	33.7 (7.5)	72.0 (7.5)
EEG, EOG, and EMG (n = 21)	67.0 (5.9)	53.1 (5.0)	41.7 (9.3)	76.3 (4.4)

SD = standard deviation, Acc = accuracy, MF1 = macro-F1 score, kappa = Cohen's kappa, AUC = area under the receiver-operating characteristic curve, EEG = electroencephalography, EOG = electrooculography, EMG = electromyography.

Impact of Removing Patients with Poor Performance

To assess the effect of poorly performing patients on the overall performance of CNN, an analysis was conducted in which patients with an initial MF1-score below 35.0% were excluded from the dataset. This resulted in the removal of seven patients from the PICU dataset. Subsequently the model was retrained on the new subset, for which EEG, EOG and EMG channels were utilised as input. The updated model performance, shown in the final row of Table 6, demonstrates a slight improvement. However, overall performance again remains low with a mean MF1 of 53.1%, suggesting that the dataset still exhibits substantial heterogeneity. Results for individual patients, corresponding to the overall results in Table 6, are provided in Appendix S.13.

Impact of Inter-Patient Training

To address the heterogeneity among patients, an inter-patient training approach was implemented to evaluate whether this improves performance. In this approach, the previously trained CNN within each cross-validation fold was utilised, and the final convolutional layer was retrained using the first two or four hours of data from each individual patient (equivalent to 240 and 480 epochs, respectively). The model was then validated on the remaining data for that patient. The results are presented in Table 7, with individual patient results available in Appendix S.14. An increase in mean accuracy from 59.1% to 73.5% and in mean MF1 from 45.9% to 51.2% is observed when comparing no inter-patient training to inter-patient training on 4 hours. A paired t-test comparing patient-specific MF1 scores yields a statistically significant p-value of 0.0016, indicating that inter-patient training can improve overall performance. However, with a mean MF1 of 51.2% and a mean kappa of 38.0% the performance again remains low.

Table 7: Performance metrics for three-state sleep stage classification for the PICU dataset comparing with and without inter-patient training. Results were obtained by 5-fold cross-validation with and without subsequently inter-patient training of 240 or 480 epochs.

	Mean three-state performance in % (SD)			
Training method	**Acc**	**MF1**	**Kappa**	**AUC**
No inter-patient training	59.1 (5.0)	45.9 (9.1)	26.5 (10.9)	69.3 (10.3)
Inter-patient training with 240 epoch (2 hours)	71.3 (9.7)	46.0 (6.1)	31.3 (7.0)	72.9 (6.9)
Inter-patient training with 480 epoch (4 hours)	73.5 (10.2)	51.2 (8.4)	38.0 (10.1)	74.1 (8.0)

SD = standard deviation, Acc = accuracy, MF1 = macro-F1 score, kappa = Cohen's kappa, AUC = area under the receiver-operating characteristic curve.

Chapter 2: Unsupervised Machine Learning Model

2.1 Research Question

Does an unsupervised machine learning model identify clear structures within sleep-EEG data from critically ill children, and do these structures either align with the conventional sleep stages or give an indication for an alternative approach for sleep staging in this population?

The results presented in Chapter 1 demonstrate that DL models and dynamic machine learning approaches achieve strong performance in sleep staging for adults and relatively good performance for non-critically ill children. However, their performance deteriorates significantly when applied to critically ill children. This highlights that conventional sleep stages are difficult to generalise to the PICU data, raising the question whether this method accurately describes sleep structures within this population. This raises the next fundamental question: Do identifiable structures, beyond the conventional sleep stages, exist in the sleep EEG data of critically ill children, or is the poor performance due to a lack of clear structure and homogeneity in the data?

To explore whether any clear structures exist across the three datasets, an unsupervised HMM will be employed. Unlike previous models that rely on manually assigned labels, this HMM will be provided with only the number of clusters to fit, without any prior knowledge of the manually assigned labels. These identified clusters will represent distinguishable structures within the data.

This approach consists of three steps. First, it will be assessed whether the model consistently produces the same clusters when trained multiple times, determining the stability of the clusters. Subsequently, the likelihood is assessed to evaluate how separable and well-defined the clusters are. Finally, these clusters will be compared with the manually scored sleep stages to determine whether they align with conventional sleep staging.

The unsupervised HMM will be tested using two approaches based on the types of input. The first approach uses CNN-extracted features, which have been trained in a supervised manner and may therefore carry an inherent bias towards the manually assigned labels (defined as the partly unsupervised approach). The second approach relies on manually selected features, ensuring a fully unsupervised method that allows the model to learn purely from the raw data, without any influence from predefined labels (defined as the fully unsupervised approach).

For healthy adults, and to some extent for non-critically ill children, it is expected that the unsupervised HMM will form stable and likely clusters that align with the conventional sleep stages, illustrating that conventional sleep stages generalise well to these populations. However, for critically ill children, the clusters are unlikely to align well with conventional labels, as the supervised models have struggled with classification in this population. However, even if the clusters do not resemble the conventional labels, the hope is that stable and likely clusters will emerge. This would suggest that there are underlying structures in the data that may require an alternative classification approach to accurately represent the sleep structure of this population.

2.2 Methods

2.2.1 Model Development

Following the same model architecture as the supervised HMM described previously (Figure 6), an unsupervised Gaussian HMM was developed. However, unlike the supervised approach, where the hidden states correspond to predefined sleep stages, the hidden states in this model represent clusters identified during the unsupervised learning process. For the observed states, two types of feature representations were used: the output of the final CNN layer before the softmax activation (partly unsupervised approach) and manually selected features (fully unsupervised approach). For clarification, the pipeline corresponding to these approaches is shown in Figure 8. The manually selected features are described by Hiemstra et al. [31] and a detailed list is provided in Appendix S.15. Normalisation and PCA was applied to both feature sets. Based on the findings from using different numbers of principal components in the supervised HMM, the partly unsupervised approach utilised the 50 first principal components. The fully unsupervised approach used 20 principal components, achieving an explained variance above 95% in all datasets. The explained variance and a visualisation of the first two principal components for both approaches are provided in Appendices S.10 and S.16, respectively. As with the supervised HMM, the emission probabilities were modelled using Gaussian distributions. The distribution of the first principal components per sleep stage for each dataset are visualised in Appendix S.4 and S.17 to assess the suitability of Gaussian emission assumptions.

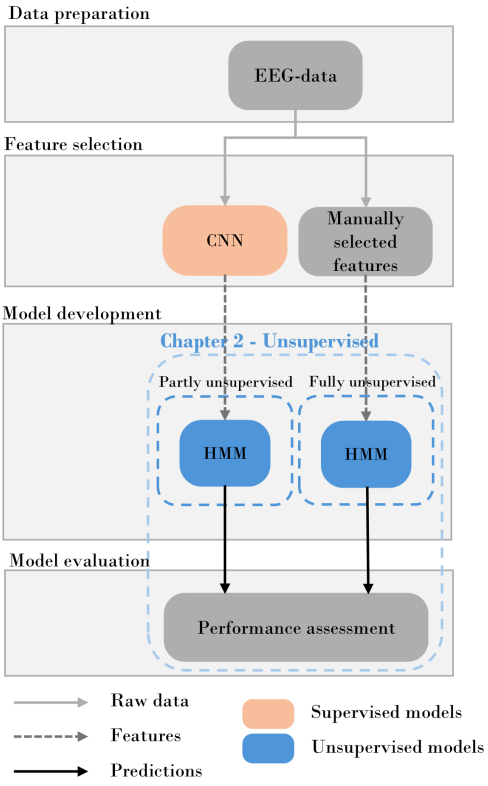


Figure 8: Overview of the pipeline of chapter 2. The figure illustrates the flow from data preparation to model evaluation for the partly and fully unsupervised approach. CNN = convolutional neural network, HMM = hidden Markov model.

2.2.2 Model Training

Using the training dataset, the emission and transition probabilities of the unsupervised HMM were learned using the Baum-Welch algorithm, which iteratively estimates and updates these probabilities to maximise the log-likelihood of the observed sequence. The log-likelihood serves as a measure of how well the model fits the observed data, providing insight into the quality of the identified clusters. As in the supervised HMM, the Viterbi algorithm was applied to determine the optimal sequence of hidden states given the observed data. Before training, the number of hidden states (i.e. clusters) to be identified by the model was specified. A detailed explanation of the algorithms is provided in Appendix S.5.

2.2.3 Model Evaluation

The same k-fold cross-validation procedure used for the supervised models was applied to the unsupervised HMM. For the Sleep-EDF dataset, k was set to 20, while for the PSG and PICU datasets, k was set to 5. In each fold, the model was trained 100 times. From every group of 10 trained models, the model with the highest log-likelihood on the validation set was saved and used to assess the stability of the model.

Once the stability evaluation was completed, the model with the highest log-likelihood from the full set of 100 trained models was selected for further evaluation.

To assess the alignment of the identified clusters with manually assigned labels, an algorithm was applied to optimally relabel the unsupervised clusters. This algorithm systematically iterates through all possible cluster-to-stage mappings and selects the mapping that maximises the MF1 score. By doing so, the most representative correspondence between the unsupervised clusters and the manually labelled sleep stages is established.

2.2.4 Performance Metrics

The stability of the unsupervised HMM was assessed using the Adjusted Rand Index (ARI), which quantifies the agreement between different cluster assignments while accounting for chance-level agreement. An ARI score of 1 indicates perfect agreement, whereas a score of 0 corresponds to random clustering. Higher ARI values therefore reflect greater consistency and stability in the identified clusters across multiple training runs. The ARI score was calculated by comparing the best model from each subset of 10 runs. This resulted in a total of 45 ARI scores per fold, from which, for all folds together, the median and percentiles were computed to provide a comprehensive measure of clustering stability.

Subsequently, the log-likelihood was assessed to determine the validity of the clustering. As the log-likelihood is influenced by both the amount of data and the number of clusters, it cannot be used to determine whether a certain value is sufficient, nor is direct comparison between datasets meaningful. Therefore, the log-likelihood of the best-performing model from the 100 training runs was compared to three reference log-likelihoods. The first reference was an HMM model trained on a Gaussian distribution, which serves as the gold standard representing the best possible score. The second reference was an HMM trained on a random distribution of the principal components, representing the worst-case scenario. This was done by randomly shuffling the values of the principal components across all sleep epochs. The third reference was the log-likelihood of the supervised HMM from chapter 1, providing a direct comparison to a model trained with predefined sleep stage labels.

Following this, the extent to which the discovered clusters align with conventional sleep stages was assessed. Following the relabelling algorithm, the classification performance was evaluated using the accuracy and macro-F1 score. For both performance metrics, the mean and standard deviation across all folds were reported. Detailed calculations of all metrics are provided in Appendix S.6.

2.3 Results - Partly Unsupervised Approach

In this section, the principal components of the features originating from the CNN are utilised as input of the unsupervised HMM.

2.3.1 Stability of the Clustering

Figure 9 presents a boxplot of the ARI scores obtained for each dataset when comparing the results of the 10 best-performing training runs of the HMM when classifying three clusters. The results indicate that the Sleep-EDF dataset exhibits the highest median ARI score and the narrowest interquartile range, suggesting more stable cluster assignments. In contrast, the PICU dataset shows the lowest median score and the widest percentile range, indicating greater variability and less stable clustering in this population.

Table 8 summarises the number of clusters identified by the HMM, for the same 10 training runs within each fold. Notably, in 36% of cases within the PICU dataset, the model identified only two clusters

instead of three. This reduction in the number of clusters increases the likelihood of achieving a higher ARI score, as fewer clusters inherently result in greater agreement across different model runs.

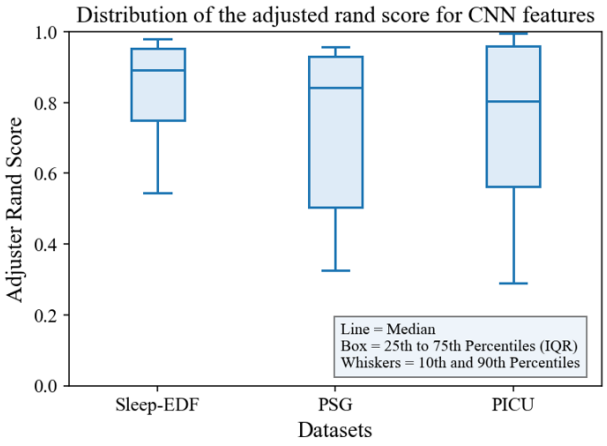


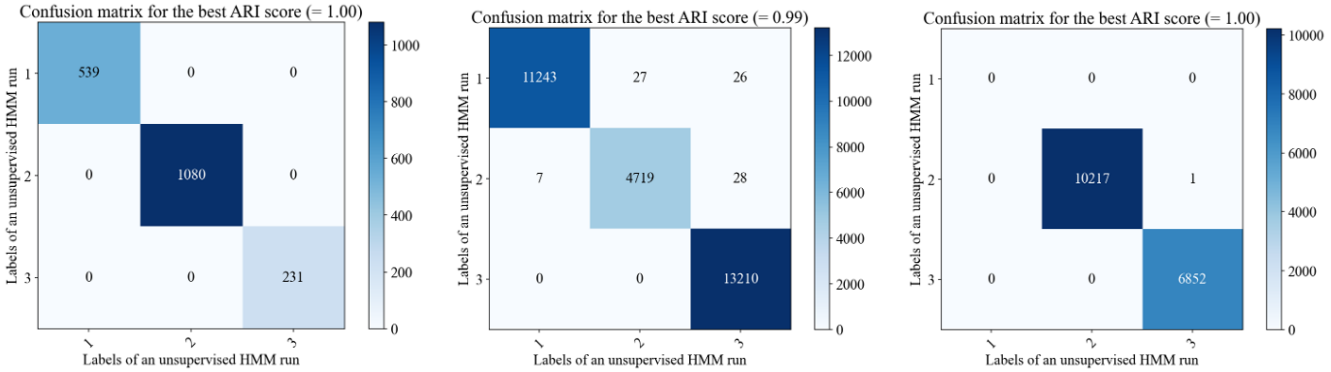
Figure 9: Boxplot comparing the clustering stability for the Sleep-EDF, PSG and PICU datasets. Adjusted rand index scores were computed between the 10 best-performing hidden Markov models when classifying three clusters, utilising features from the convolutional neural network. The mean and percentiles were calculated over all 20 folds for the Sleep-EDF dataset and over all 5 folds for the PSG and PICU datasets. CNN = convolutional neural network, IQR = inter quartile range.

Table 8: Number of clusters identified for each training run of the hidden Markov model for the Sleep-EDF, PSG and PICU dataset. The number of clusters identified by the 10 best-performing models when classifying three clusters are shown. Results were obtained by summation over all 20 folds for the Sleep-EDF dataset and over all 5 folds for the PSG and PICU dataset.

	Number (%) of clusters		
Dataset	**One**	**Two**	**Three**
Sleep-EDF	0 (0)	10 (5)	190 (95)
PSG	0 (0)	0 (0)	50 (100)
PICU	0 (0)	18 (36)	32 (64)

Figure 10 presents confusion matrices for each dataset, comparing the results of two individual runs. These matrices correspond to the models that achieved the highest ARI scores when compared against each other. The cluster numbers have been relabelled to ensure alignment between the two models, allowing the clusters to be visualised along the diagonal. The results show that the highest ARI scores are achieved when three clusters are identified in the Sleep-EDF and PSG datasets. In contrast, for the PICU dataset, the best agreement is observed when only two clusters are identified, suggesting that it is more difficult to identify three stable clusters in this dataset.

Confusion matrices for models that achieved the lowest ARI scores can be found in Appendix S.18.



(a) Sleep-EDF Dataset **(b)** PSG Dataset **(c)** PICU Dataset

Figure 10: Confusion matrices showing the best alignment between partly unsupervised clusterings of two hidden Markov models. The results are shown for the models that achieved the highest adjusted rand score when compared to each other. Features from the convolutional neural network were utilised. Cluster numbers were relabelled to match each other, visualising the clusters on the diagonal. ARI = adjusted rand score.

2.3.2 Likelihood of the Clustering

Table 9 presents the log-likelihoods obtained when training the unsupervised HMM for each dataset, providing a comparison against best- and worst-case scenarios and the supervised HMM. All models were trained to classify three clusters and the best-performing model of 100 training runs was utilised. To improve interpretability, all log-likelihood values have been scaled by multiplying them by -10^5, where a lower log-likelihood corresponds to a more probable clustering. As previously mentioned, direct comparisons between datasets are not meaningful.

For all datasets, the best log-likelihood values are consistently achieved when using Gaussian-distributed features, with mean values falling outside the standard deviation range of the other results, which suggests a significantly better identification of three clusters. For the Sleep-EDF and PICU datasets, the log-likelihoods for the supervised HMM, shuffled features, and actual features all fall within each other's standard deviations. Consequently, no definitive conclusion can be drawn regarding whether any of these approaches produce a more probable clustering than the others, indicating that the clustering is no more likely than clustering performed on random features. In the PSG dataset, however, the likelihood values for the supervised HMM and the unsupervised HMM using actual features fall outside the range of the HMM using shuffled features. This suggests that clustering within this dataset is more probable than random clustering.

Table 9: Log-likelihoods for classification of three clusters by a hidden Markov model utilising features from the convolutional neural network. The results were obtained from the best-performing model within each fold of the 20 or 5-fold cross-validation for the Sleep-EDF and PSG/PICU dataset, respectively . All log-likelihoods were multiplied by -10^5 for interpretability. A lower log-likelihood indicates more likely clustering. Comparison between datasets is not valid.

	Mean (SD) log-likelihood for three clustering			
Dataset	**Supervised HMM**	**Unsupervised HMM**		
		Shuffled features (Worst scenario)	**Gaussian features (Best scenario)**	**Actual features (study's scenario)**
Sleep-EDF (n = 39)	27.7 (5.6)	29.4 (5.5)	11.3 (2.1)	27.9 (5.4)
PSG (n = 120)	379.0 (11.6)	417.8 (10.7)	157.4 (5.8)	380.1 (19.8)
PICU (n = 28)	204.4 (36.3)	215.0 (34.7)	79.8 (11.4)	233.1 (45.1)

SD = standard deviation, HMM = hidden Markov model

2.3.3 Comparison of the Unsupervised Clusters to Conventional Sleep Stages

To assess whether the clusters identified by the unsupervised HMM align with manually labelled sleep stages, the clusters are relabelled to best match these labels. This relabelling process is applied across different numbers of unsupervised clusters (3 to 10) and is done to match three-state classification (wake, NREM and REM), of which the performance is presented in Table 10. The findings from the Sleep-EDF and PSG datasets indicate that performance generally improved as the number of clusters increased, for which multiple clusters can fall within one sleep stage. A significant increase in the MF1-score is observed up to five or six clusters, beyond which improvements became marginal.

For the Sleep-EDF dataset, a maximum MF1 of 73.6% was achieved, indicating that the identified clusters reasonably align to the manually scored sleep stages. For the PSG dataset, a somewhat lower maximum MF1 of 65.5% is observed. For the PICU dataset, a maximum MF1 of only 40.1% is reached, suggesting that clusters can not accurately be relabelled to align with manually scored labels. Interestingly, the maximum MF1 score observed for the PSG and PICU dataset are higher than achieved by the supervised HMM-f (Table 3), indicating a possible contribution of defining a sleep stage through multiple clusters.

Table 10: Performance metrics of the unsupervised hidden Markov model with varying number of clusters after relabelling for three-state classification on the Sleep-EDF, PSG and PICU dataset. The results were obtained by 20- or 5-fold cross-validation for the Sleep-EDF and PSG/PICU dataset, respectively. The performance is calculated after relabelling the unsupervised clusters to find the highest macro-F1 score when compared to manually scored labels.

		Number of clusters classified							
Dataset	**Mean (SD) metric in %**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Sleep-EDF	Acc	58.6 (9.8)	71.7 (10.8)	78.0 (8.0)	76.7 (8.8)	76.9 (9.1)	78.3 (5.1)	77.9 (8.8)	78.1 (8.5)
(n = 39)	MF1	52.7 (8.5)	65.1 (11.5)	71.6 (9.8)	71.2 (8.8)	70.1 (11.7)	71.8 (8.0)	73.0 (9.2)	73.6 (8.9)
PSG	Acc	54.6 (3.4)	44.0 (3.4)	60.7 (5.0)	63.2 (4.5)	67.8 (4.8)	68.1 (5.6)	70.1 (1.9)	64.6 (1.4)
(n = 120)	MF1	51.7 (5.3)	35.5 (3.8)	52.0 (5.2)	60.5 (4.9)	61.7 (5.8)	63.8 (5.3)	65.5 (2.7)	56.5 (2.6)
PICU	Acc	50.2 (5.4)	53.9 (11.4)	56.2 (10.8)	55.6 (10.5)	57.7 (10.0)	54.2 (12.2)	56.1 (10.8)	55.8 (11.1)
(n = 28)	MF1	37.0 (4.7)	36.3 (3.9)	39.2 (4.3)	38.0 (3.5)	40.1 (6.9)	39.4 (4.9)	39.4 (4.4)	39.5 (3.9)

SD = standard deviation, Acc = accuracy, MF1 = macro F1-score.

When varying the number of clusters, the following question arises: do the clusters still remain stable? To investigate this, Appendix S.19 presents a boxplot for each dataset, which visualises ARI scores for varying amount of clusters to classify. The results show that as the number of clusters increases, the median ARI score decreases, indicating reduced cluster stability. In the same appendix, the number of actually identified clusters is presented for different amount of clusters as input. Across all datasets, it is observed that as the number of input clusters increases, this maximum number of clusters is less frequently identified. This indicates that there is a maximum to the distinct clusters that can be identified.

2.4 Results - Fully Unsupervised Approach

In this section, the principal components of the manually selected features are utilised as input of the unsupervised HMM, such that no supervised information is used for the clustering at all.

2.4.1 Stability of the Clustering

Figure 11 presents a boxplot of the ARI scores obtained for each dataset when comparing the results of the 10 best-performing training runs when classifying three clusters. Compared to the boxplot in Figure 9, the medians are lower and the ranges are wider for the Sleep-EDF and PSG datasets, indicating less stable clusters. In contrast, the median ARI for the PICU dataset is very high, with ranges spanning from 0 to 1. This variability can be explained by the results in Table 11, which shows the number of clusters classified in each training run. In 18% of the runs, only a single cluster was identified, automatically yielding an ARI score of 1 when compared to other runs classifying one cluster. However, when comparing these runs to runs that assigned multiple clusters, very low ARI scores will be calculated, explaining the wide range. For the Sleep-EDF and PSG datasets, three clusters were consistently identified.

Figure 12 presents confusion matrices for the two runs achieving the highest ARI score when compared. Cluster numbers were again relabelled to visualise clusters along the diagonal. In both the Sleep-EDF and PSG datasets, three clusters are identified; however, cluster sizes vary significantly in the PSG dataset. In contrast, only a single cluster is identified in the PICU dataset, resulting in a perfect ARI score.

Confusion matrices for the models that achieved the lowest ARI scores can be found in Appendix S.18.

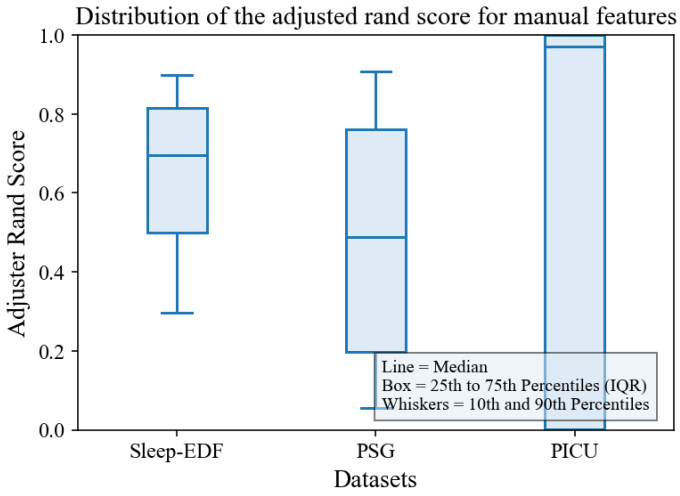
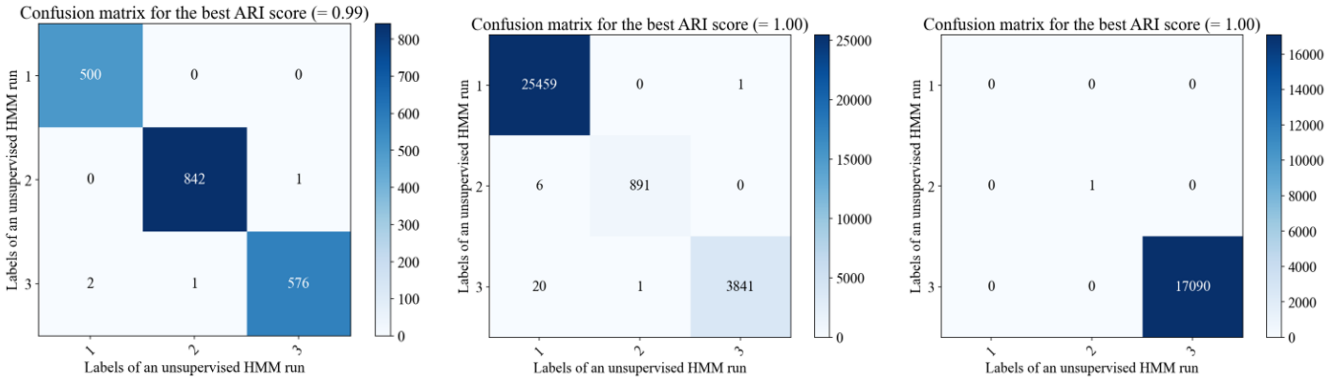


Figure 11: Boxplot comparing the clustering stability for the Sleep-EDF, PSG and PICU datasets. Adjusted rand index scores were computed between the 10 best-performing hidden Markov models when classifying three clusters, utilising manually selected features. The mean and percentiles were calculated over all 20 folds for the Sleep-EDF dataset and over all 5 folds for the PSG and PICU datasets. IQR = inter quartile range.

Table 11: Number of clusters identified for each training run of the hidden Markov model for the Sleep-EDF, PSG and PICU dataset. The number of clusters identified by the 10 best-performing models when classifying three clusters are shown. Results were obtained by summation over all 20 folds for the Sleep-EDF dataset and over all 5 folds for the PSG and PICU dataset.

	Number (%) of clusters		
Dataset	**One**	**Two**	**Three**
Sleep-EDF	0 (0)	0 (0)	200 (100)
PSG	0 (0)	0 (0)	50 (100)
PICU	9 (18)	34 (68)	7 (14)



(a) Sleep-EDF Dataset **(b)** PSG Dataset **(c)** PICU Dataset

Figure 12: Confusion matrices showing the best alignment between partly unsupervised clusterings of two hidden Markov models. The results are shown for the models that achieved the highest adjusted rand score when compared to each other. Manually selected features were utilised. Cluster numbers were relabelled to match each other, visualising the clusters on the diagonal. ARI = adjusted rand score.

2.4.2 Likelihood of the Clustering

Table 12 presents the log-likelihoods obtained when training an HMM on manual features for different scenarios for each dataset. As before, all log-likelihood values have been multiplied by -10^5, with lower values indicating a higher likelihood. For the Sleep-EDF dataset, only small differences are observed between log-likelihood values, with most falling within each other's standard deviation range. This makes it difficult to interpret the likelihood of the actual scenario. For the PSG dataset it stands out that the supervised HMM achieves the best log-likelihood. Thereby, the worst-case scenario yields a better log-likelihood than the actual features, suggesting the absence of a clear structure in the actual scenario. In the PICU dataset it again stands out that most log-likelihoods fall within each others SD range. However, the actual scenario shows the worst log-likelihood, again suggesting a lack of clear patterns in the data.

Table 12: Log-likelihoods for classification of three clusters by a hidden Markov model utilising manually selected features. The results were obtained from the best-performing model within each fold of the 20 or 5-fold cross-validation for the Sleep-EDF and PSG/PICU dataset, respectively . All log-likelihoods were multiplied by -10^5 for interpretability. A lower log-likelihood indicates more likely clustering. Comparison between datasets is not valid.

	Mean (SD) log-likelihood for three clustering			
Dataset	**Supervised HMM**	**Unsupervised HMM**		
		Shuffled Features (Worst Scenario)	**Gaussian Features (Best Scenario)**	**Actual Features (Study's Scenario)**
Sleep-EDF (n = 39)	5.3 (0.9)	5.9 (0.9)	4.7 (0.6)	5.6 (1.0)
PSG (n = 120)	61.3 (3.9)	77.0 (5.3)	65.4 (3.3)	84.6 (29.7)
PICU (n = 28)	36.7 (9.6)	38.8 (8.6)	32.6 (4.7)	53.1 (8.7)

SD = standard deviation, HMM = hidden Markov model

To illustrate the performance of the supervised HMM on manual features, results are provided in Appendix S.20. For three-state classification, MF1-scores of 76.4%, 54.0%, and 37.3% are achieved for the Sleep-EDF, PSG, and PICU datasets, respectively. These findings indicate that manual features can yield relatively strong results for healthy adults, but performance declines significantly in non-critically ill and critically ill children.

2.5 Analysis of the Results

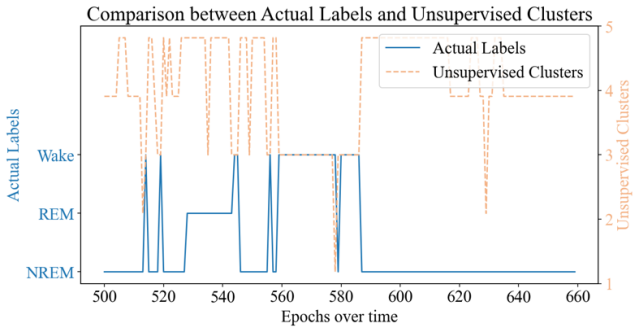
To further analyse the results from the unsupervised HMM on patient level, three methods are employed. The first method involves plotting a two hypnograms over time: one representing the sleep stages as manually labelled (Wake, NREM, and REM), and the other showing the unsupervised clusters. For all visualisations, the HMM is trained to identify five clusters. This approach illustrates whether the HMM identifies stable clusters over time and how well these clusters align with the conventional sleep stages.

The second method visualises the probabilities of each cluster over time. Each value ranges from zero to one, where one indicates high certainty that an epoch belongs to a given cluster, and zero suggests it is highly unlikely. The model shows high confidence when values are close to zero or one, while lower confidence is reflected by values falling in between.

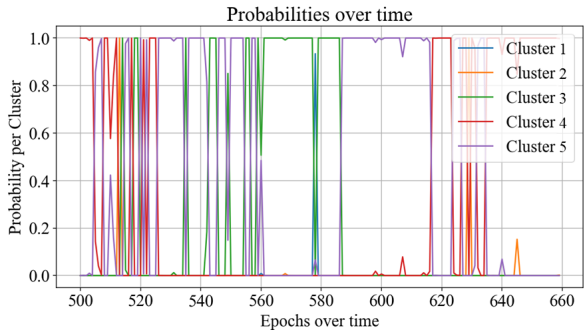
The third method presents a visualisation of the first two principal components. This is plotted twice: in the first visualisation, points are coloured according to the actual sleep stage labels, while in the second, they are coloured based on the clusters assigned by the unsupervised HMM. This allows for a direct visual comparison between the clusters identified by the model and the conventional sleep stage labels.

2.5.1 Comparison between Partly and Fully Unsupervised Approaches

In Figure 13, a hypnogram and probability plot are presented for a selected time period from a patient in the Sleep-EDF dataset, to visualise clusters classified by the partly unsupervised approach. In Figure 13a, it can be observed that during a wake period, the HMM classifies a stable cluster. However, during REM and NREM, the unsupervised clusters vary more, with for example cluster 5 appearing in both stages. When examining Figure 13b, it is evident that when a cluster is classified for an extended period, it is done with high certainty (i.e. with a probability score of one). Additionally, many transitions between clusters occur with strong confidence, as the probabilities shift rapidly between zero and one. However, particularly at the beginning, there are periods where cluster assignments are made with lower certainty.



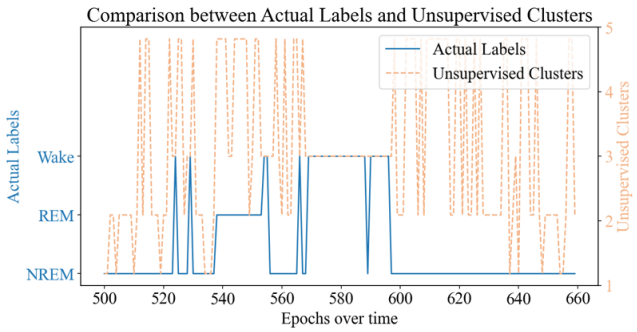
(a) Hypnograms over time. For comparison between manually labelled sleep stages to partly unsupervised clusters.



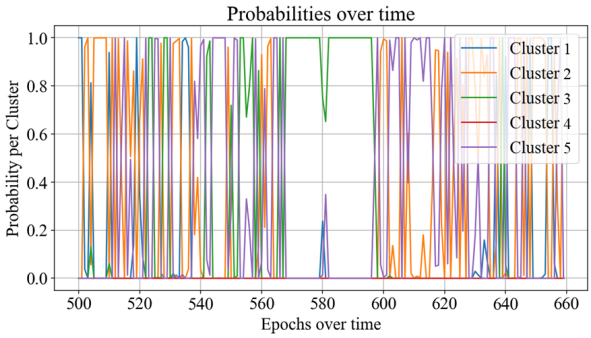
(b) Probabilities over time. A higher probability indicates a greater likelihood of an epoch to belong to that cluster.

Figure 13: Visualisation of partly unsupervised clusters for a selected time period of a patient from the Sleep-EDF dataset. Clusters (a) and their corresponding probabilities (b) were obtained from an unsupervised hidden Markov model utilising features from the convolutional neural network.

In Figure 14, the same time period from the same patient is shown, this time with unsupervised clusters derived using the fully unsupervised approach. As illustrated in Figure 14a, wakefulness again corresponds to a stable cluster. In contrast, during REM and NREM sleep, the cluster assignments fluctuate more frequently. Additionally, Figure 14b reveals that, apart from the wake period, the model's confidence in cluster assignments is generally lower, reflected by a greater number of probability values falling between zero and one. This suggests less distinct clustering during sleep.



(a) Hypnograms over time. For comparison between manually labelled sleep stages to partly unsupervised clusters.



(b) Probabilities over time. A higher probability indicates a greater likelihood of an epoch to belong to that cluster.

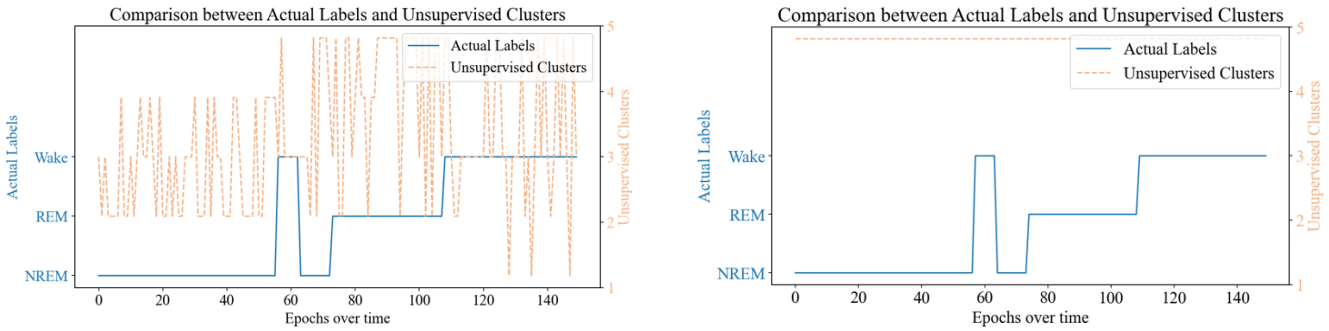
Figure 14: Visualisation of fully unsupervised clusters for a selected time period of a patient from the Sleep-EDF dataset. Clusters (a) and their corresponding probabilities (b) were obtained from an unsupervised hidden Markov model utilising manually selected features.

These findings suggest that, aside from periods of wakefulness, the unsupervised approaches do not identify stable clusters over time. This indicates that sleep scoring using an HMM based solely on data, thus without incorporating criteria defined by the AASM, is challenging, even in a healthy adults.

Hypnograms and probability plots for an entire night's sleep from this patient can be found in Appendix S.21, along with visualisations of the principal components. Results for the PSG dataset are provided in the same appendix, which show similar patterns to those observed in the Sleep-EDF dataset.

2.5.2 Analysis of the PICU dataset

Figures 15a and 15b present hypnograms over a time period from a patient in the PICU dataset, where left shows partly unsupervised clusters and rights shows fully unsupervised clusters. For the partly unsupervised approach, the clusters fluctuate considerably, with no stable clusters visible. Even during wakefulness, the model fails to identify a stable cluster, suggesting that the clusters do not align well with the conventional sleep stages. In contrast, for the fully unsupervised approach, all epochs are assigned to a single cluster throughout the entire period, indicating that no distinct structures are found.

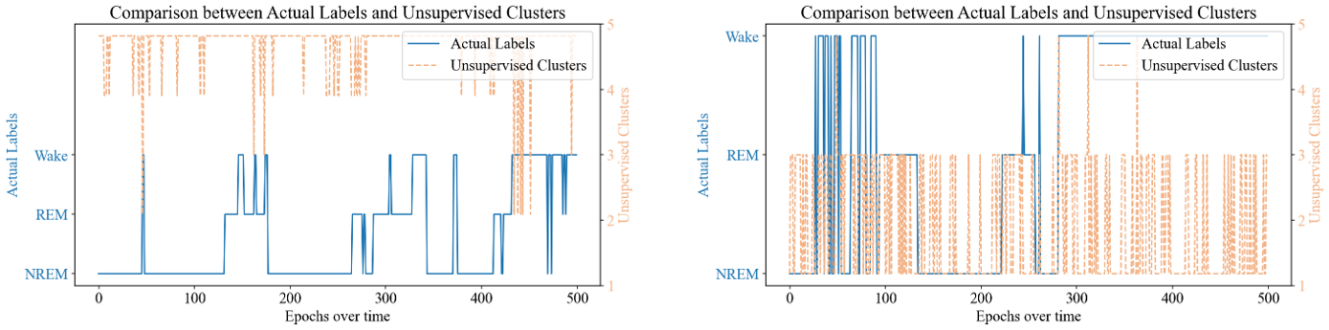


(a) Partly unsupervised approach

(b) Fully unsupervised approach

Figure 15: Hypnograms over time for comparison between manually labelled sleep stages and unsupervised clusters for a selected time period of a patient from the PICU datast. Clusters were obtained from an unsupervised hidden Markov model utilising features from the convolutional neural network (a) or manually selected features (b).

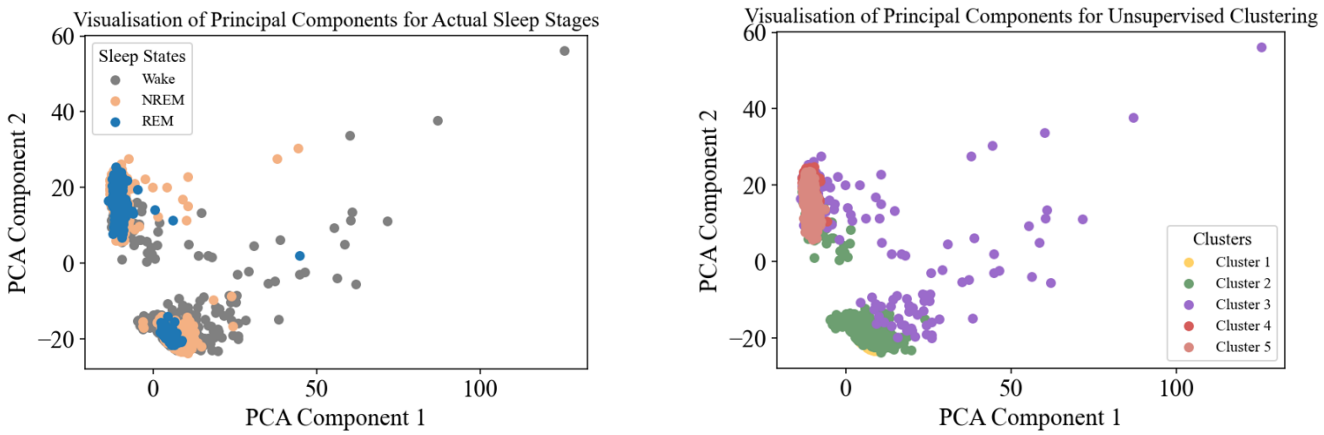
Figure 16 presents hypnograms from two different PICU patients within the same validation fold. Both clusterings were obtained through the partly unsupervised approach. For the patient in Figure 16a, most epochs are assigned to unsupervised clusters 4 and 5, whereas for the patient in Figure 16b, all epochs are classified as cluster 1 and 2. This is also visualised in Figure 17, where for these two patients the first two principal components are observed. Two distinct clusters can be visualised, both presenting an individual patient. This suggests that the differences between patients' EEG data are so substantial that the model is more likely to assign different clusters to different patients rather than distinguishing between different sleep structures.



(a) PICU patient 1

(b) PICU patient 2

Figure 16: Hypnograms over time for comparison of unsupervised clusters between two patients from the PICU dataset. Clusters were obtained from an unsupervised hidden Markov model through the partly unsupervised approach. The PICU patients originate from the same cross-validation fold.



(a) Actual sleep stages

(b) Unsupervised clusters

Figure 17: Comparison between manually labelled sleep stages and partly unsupervised clusters through visualisation of the first two principal components for two patients from the PICU dataset. Clusters were obtained from an unsupervised hidden Markov model utilising features from the convolutional neural network. Two distinct clusters, resembling the two patients, are visible.

These findings suggest that, for the PICU dataset, neither the partly nor the fully unsupervised approach is capable of identifying clusters within the data. In particular, the clusters lack stability over time and are not generalisable across patients. This indicates that an alternative approach to sleep staging, beyong the conventional sleep stages, is not feasible through utilisation on an unsupervised HMM.

Discussion

This study examined whether DL-based feature extraction and dynamic machine learning models can accurately classify sleep stages in healthy adults, non-critically ill and critically ill children, using single-channel EEG data. To further gain insight into this data, an unsupervised model was applied to assess whether clear structures could be identified without reliance on manually labelled data.

All models were evaluated on three datasets: the Sleep-EDF dataset of healthy adults (used as a benchmark and for comparison with previous studies), the PSG dataset of non-critically ill children (to assess the impact of age on sleep staging), and the PICU dataset, comprising EEG recordings from critically ill children. The supervised models performed well on the Sleep-EDF dataset, fairly well on the PSG dataset, and poorly on the PICU dataset. These findings indicate that, while DL methods can effectively extract relevant features for sleep staging in healthy populations and dynamic models can occasionally enhance performance, these approaches are unable to compensate for the substantial variability and atypical sleep architecture observed in critically ill children.

The results obtained from the PICU dataset, supported by existing literature describing the distinct characteristics of sleep in both children and critically ill patients, shifted the focus towards identifying whether other underlying structures exist, without relying on the conventional sleep stages [15, 16]. When applying an unsupervised HMM, some discernible structures were observed in the Sleep-EDF and PSG datasets. However, the model mostly struggled to form distinct and temporally stable clusters for both the partly and fully unsupervised approaches. In the PICU dataset, most epochs within individual patients were assigned to only one or two clusters, suggesting an absence of multiple clear and separable clusters in the data. Moreover, the distribution of clusters varied considerably between patients, reflecting a high degree of inter-patient variability.

Taken together, the findings from the unsupervised approach underscore the difficulty of identifying distinct and stable clusters over time across all datasets, highlighting the challenge of performing classification based solely on EEG data. This reinforces the value of rule-based classification systems, such as those established by the AASM, which offer standardised criteria for manual scoring. These guidelines are reasonably well generalisable to populations of healthy adults and non-critically ill children, as also reflected in the performance of the supervised models. However, it is clear that both rule-based approaches and the solely data-driven alternative evaluated in this study fall short in the context of critically ill children in the PICU. For this population, neither DL-based automatic sleep staging based on conventional scoring guidelines nor unsupervised models currently offer a reliable solution for accurate classification.

Methodological constraints

To investigate whether incorporating temporal context could enhance classification performance by accounting for the cyclic nature of sleep, both an LSTM network and an HMM were implemented as dynamic models. For both models, a smoothing effect was visually observed in the predicted sleep stages, indicating that short-term dependencies between adjacent epochs were captured. However, temporal modelling did not consistently lead to improved classification performance across datasets. The most notable gain from the LSTM model was seen when applied to the full PSG dataset comprising 120 patients, indicating that large training sets may be essential for LSTMs to learn temporal patterns. Furthermore, although LSTMs are designed to capture long-term dependencies, no evidence was found that such long-term patterns were effectively learnt (i.e. no improvement in performance was observed with longer sequence lengths). In contrast, HMMs are inherently limited to modelling short-term dependencies and cannot learn long-term

temporal structures. Thus, both models primarily reflected the increased likelihood of consecutive epochs belonging to the same sleep stage, rather than capturing the broader cyclic nature of sleep-such as the typical duration of a full sleep cycle.

The CNN and LSTM architectures and hyperparameters were based on the work by Supratak et al., which utilised the Sleep-EDF dataset, and were not further fine-tuned for the PSG and PICU datasets. While further optimisation may have improved results to some extent, this was not prioritised, as the primary aim was to better understand the nature of the data rather than to maximise classification accuracy. Therefore, unsupervised approaches were further analysed. Moreover, given the low baseline performance in the PICU dataset, even optimised models were unlikely to achieve clinically meaningful performance.

Furthermore, DL approaches, such as CNNs and LSTMs, bring two important limitations. First, the features extracted by the convolutional neural network are inherently difficult to interpret, often referred to as "black box" representations. This hinders the ability to relate model outputs to known physiological markers or to understand how features differ across age groups and clinical populations. Second, DL methods are computationally intensive, requiring substantial processing time and resources, which can limit feasibility for real-time use in clinical settings.

By contrast, HMMs offer a more computationally efficient and interpretable alternative to LSTMs. When performance is comparable, their simplicity may be advantageous. However, a key limitation lies in their assumption of Gaussian-distributed input features, an assumption that does not hold for all principal components as shown in Appendix S.4 and S.17. This may partially explain the suboptimal classification performance in supervised settings and the instability of clusters in the unsupervised application. Moreover, HMMs are less suited to handling high-dimensional input, making it necessary to reduce feature dimensionality through methods such as PCA. This introduces an additional constraint: selecting the appropriate number of principal components is challenging, as too few limits the amount of explained variance, while too many are difficult for the HMM to handle effectively. As a result, the use of CNN-derived features required a trade-off that ultimately led to a relatively low proportion of explained variance.

Moreover, the unsupervised HMM was evaluated using two input strategies, each with inherent drawbacks. The approach using manually selected features introduces bias through human assumptions about which signal characteristics are informative. In contrast, the approach using CNN-extracted features is not shaped by human assumptions, but the features themselves were learned through supervised training and therefore reflect biases related to the labelled data. Thus, neither input strategy provides a fully objective or assumption-free representation of the EEG signals.

Lastly, validating the output of the unsupervised HMM is inherently challenging due to the absence of ground truth labels. While clustering stability was assessed using the ARI score, this metric does not provide insight into the relevance or quality of the clusters. The log-likelihood of the model reflects its confidence in making a good prediction, but is influenced by factors such as the number of data points and clusters, making it difficult to compare across datasets or define a clear threshold for good performance. As a result, it remains difficult to determine the clinical significance or quality of the clusters and structures identified by the unsupervised model.

Comparison with Previous Studies

The results obtained on the healthy adult Sleep-EDF dataset are in line with previous studies in the literature. For example, Supratak et al., on which the CNN and LSTM of this study are based, reported an accuracy and macro-F1 score of 82.0% and 76.9%, respectively, for five-state sleep classification using a CNN-BiLSTM architecture, which is comparable to the results observed in this study [30]. Other studies

using similar DL models also report comparable performance on this dataset, confirming the suitability of DL for sleep staging in healthy adult populations and the correct construction of models in this study [14].

For the PSG dataset containing non-critically ill children, a previous thesis that used an XGBoost classifier combined with manually engineered features reported accuracy and kappa scores of 79% and 66% for three-state classification. When externally validated on the PICU dataset, these scores dropped to 55% and 34%, respectively [31]. Although the results are similar to those obtained in the present study, direct comparisons are complicated due to differing class definitions. In the other study, sleep stages were grouped as Wake, slow-wave sleep (SWS), and non-slow-wave sleep (NSWS). SWS was defined as N3, while NSWS consisted of N1, N2, and REM. When solely wake, NREM and REM were labelled, SWS included all NREM epochs. These differences in definitions likely positively influenced the classification performance, as the two easiest to classify stages (N3 and wake) were classified separately and the rest were grouped into one label. No studies are published on automatic sleep staging on other populations of critically ill children, likely because conventional sleep stages generalise poorly to this group.

As previously described, DL has the drawbacks of high computational cost and limited interpretability, which may favour the use of simpler approaches when they achieve comparable results. However, to mitigate the uncertainty around whether manually selected features are truly the most informative, DL-based feature extraction may be preferable. This can be particularly relevant in contexts where the relevance of known features is unclear, such as in novel or complex datasets.

Limitations

Several limitations of this study must be acknowledged. Firstly, the size and quality of the PICU dataset are limited. As previously noted by Hiemstra et al., EEG recordings obtained in the PICU frequently suffer from noise, artefacts, and signal discontinuities, due to the assessment in a real-life intensive care environment. In comparison to the PSG dataset, the PICU data exhibited a higher prevalence of artefacts caused by 50-Hz electrical interference and ECG contamination, both of which can increase electrode impedance. These disturbances are likely to compromise the performance of the models [31].

In addition, the heterogeneity of the dataset, encompassing a range of patient ages, diagnoses, medications, and illness severities, adds further complexity to the learning process. An evaluation of the PSG dataset demonstrated that age significantly influences the difficulty of accurate sleep staging. However, it remains challenging to determine the specific impact of other patient characteristics on model performance. Even among patients with seemingly similar characteristics, substantial variability in classification results was observed, as shown in Appendix S.13. Furthermore, excluding patients who were difficult to classify did not lead to a substantial improvement in overall performance. These findings suggest that the reduced performance on the PICU dataset is likely multifactorial, with both compromised data quality and high variability among patients playing a substantial role.

Another key limitation relates to the manual scoring of sleep stages in critically ill children. While the AASM guidelines are well-established for use in healthy individuals, their applicability to critically ill paediatric patients is questionable. Manual scoring by clinical neurophysiology laborants is the gold standard, however, they also have difficulty accurately scoring the sleep of patients in the PICU dataset, reflected by the high amount of sleep scored as NREM. NREM is a normal stage to be scored in young children, but in this population is also often used in older children (as shown in Appendix S.12), reflecting that it was not possible to classify an epoch as N1, N2 or N3 [9]. Studies examining inter-rater variability in manual sleep staging for critically ill adults have reported Cohen's kappa values ranging from 0.52 to as low as 0.19, highlighting the substantial challenges of sleep classification in the ICU environment [42,

43]. For children, this difficulty is further increased by the frequent absence or alteration of physiological markers that are typically used to define sleep stages [15, 16]. These issues raise important concerns regarding the validity of using manually labelled data as a gold standard in this population.

Furthermore, this study relied on data from a single centre, meaning that the PSG and PICU datasets may not be representative of other clinical settings where different equipment, protocols, and patient populations are present. This limits the generalisability of both the models and the findings. While external validation on another PICU dataset may be of limited value given the low performance observed, validation on an independent dataset of non-critically ill children would be highly relevant to assess the robustness and applicability of the models in broader contexts.

Future Directions

This study demonstrates that ML approaches hold promise for supporting automatic sleep stage classification in healthy adults and non-critically ill children and the next step would be to include a clinical validation study to assess the applicability and reliability in clinical practice. These studies could evaluate the integration of ML models into clinical workflows to assist sleep technicians or neurophysiologists. For instance, models could provide preliminary sleep stage assignments for each EEG epoch, which clinicians then verify and adjust, enabling a semi-automated scoring process that reduces workload. In addition, active learning strategies, where the model asks clinicians for feedback on the most uncertain epochs, may further improve performance over time, potentially increasing the applicability for clinical practice. Subsequently, external validation is necessary to assess applicability in broader contexts.

In contrast, automatic sleep staging in critically ill children remains a major challenge. The poor classification performance of all models on the PICU dataset suggests that existing sleep stage definitions may not generalise well to this highly heterogeneous population. Therefore, further investigation of sleep staging with conventional labels, such as through the use of more complex models, is not recommended when relying solely on the current PICU dataset.

Thus, to improve the understanding of sleep characteristics in critically ill children, one possible approach is to utilise a larger dataset that allows for the formation of more homogeneous subsets, thereby accounting for the high degree of heterogeneity within this population. However, in the current study, no clear evidence was found that specific patient characteristics, such as age, diagnosis, or medication, consistently influenced classification performance. As a result, it remains unclear how such subgroups should be defined and whether sufficient homogeneity can realistically be achieved within them. Therefore, a more extensive dataset is needed to gain more insight into the patient characteristics and to enable meaningful stratification of the population, after which new ML-approaches can be attempted.

However, given the multifactorial nature of sleep disruption in critically ill children, where multiple patient characteristics may interact, manual subgrouping is likely to be challenging. As an alternative, it may be more effective to incorporate these clinical variables directly into the model as additional input features. This approach enables the model to take patient-specific context into account during training and prediction, allowing it to adjust its interpretation of EEG patterns based on relevant clinical background. By doing so, the model possibly captures the heterogeneity of the population better without requiring predefined subgroups.

In addition to patient characteristics, the effect of signal noise and artefacts should also be further examined in a larger dataset. This may enable the identification of thresholds or criteria for acceptable signal quality, which could help with data inclusion. By combining stratification, integration of clinical variables, and analysis of signal quality within a more comprehensive dataset, future work may uncover

more robust structures in the data and clarify for which patient groups accurate sleep staging is feasible.

If the goal is to improve classification of the PICU data using conventional sleep stages, domain adaptation techniques could be explored. These methods aim to transfer knowledge from a well-annotated source domain, such as Sleep-EDF or the PSG dataset, to a more complex target domain, like the PICU population. For example, a model trained on the PSG dataset could be adapted to account for differences in signal characteristics or stage distribution in the PICU dataset. Domain adaptation can be applied at multiple levels: at the input level, by transforming EEG features from the target domain to better match the source domain; and at the output level, by aligning label distributions to reflect how sleep stages appear in the PICU population. However, successful domain adaptation requires both a well-performing source model and sufficient data from the target population to enable adaptation, again reinforcing the need for a larger and more comprehensive PICU dataset.

If the aim is instead to discover novel patterns in the EEG data, beyond the scope of conventional staging, a possibly promising direction lies in the use of DL-based unsupervised methods, such as autoencoders. Autoencoders reduce dimensionality similarly to PCA, but can capture complex, non-linear relationships and denoise signals, potentially revealing more useful features. When combined with a clustering method, these representations could uncover clear structures in the EEG that are not evident using simpler methods. Again, this approach requires a larger dataset to distinguish patterns from inter-patient variability.

To support either of these directions, future work should place greater emphasis on model explainability, which remains a key challenge in DL. Techniques such as attention mechanisms can help identify which features or time segments most influence the model's predictions. These insights may improve understanding of the neurophysiological mechanisms of sleep, particularly in complex populations. Ultimately, they may guide the development of simpler, interpretable models based on dataset-specific features.

Lastly, incorporating additional physiological signals such as ECG or respiration could offer further improvements. These signals are easier to acquire, less susceptible to neurological variability, and have shown potential in adult sleep staging [44, 45]. Although their use in paediatric and critically ill populations is less established, previous work using ECG-derived features on the PSG dataset achieved balanced accuracies around 59–61% for three-stage classification [46]. While such signals are unlikely to fully address the challenges of the PICU dataset, they may help improve model performance across different populations.

Conclusion

This study demonstrated the potential of DL-based feature extraction and dynamic models for automatic sleep stage classification in healthy adults and non-critically ill children using single-channel EEG. In contrast, performance in critically ill children was substantially lower. This raises important questions about the generalisability of conventional sleep stage definitions to this highly heterogeneous population, where a complex combination of contributing factors likely drives the reduced performance. Moreover, results across all datasets suggest that using a purely data-driven approach with an unsupervised HMM does not reliably identify clear alternative structures beyond the conventional sleep stages.

At present, automatic sleep staging in critically ill children remains a significant challenge, and clinical decisions should not rely on model outputs without further research. Additional analyses, whether following the methodology of this study or employing more advanced approaches, are needed within a larger dataset to improve the understanding of sleep in this population. Nevertheless, in healthy adults and non-critically ill children, machine learning models show promise for supporting (semi-)automated sleep staging, potentially enhancing the efficiency of clinical workflows for manual sleep scoring.

List of Supplementary Materials

References

1. Luyster FS, Strollo PJ, Zee PC, and Walsh JK. Sleep: A Health Imperative. en. Sleep 2012 Jun; 35:727–34. DOI: 10.5665/sleep.1846. Available from: https://academic.oup.com/sleep/article-lookup/doi/10.5665/sleep.1846 [Accessed on: 2024 Sep 20]
2. Kudchadkar SR, Aljohani OA, and Punjabi NM. Sleep of critically ill children in the pediatric intensive care unit: A systematic review. en. Sleep Medicine Reviews 2014 Apr; 18:103–10. DOI: 10.1016/j.smrv.2013.02.002. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1087079213000257 [Accessed on: 2025 Jan 10]
3. Carno MA and Connolly HV. Sleep and Sedation in the Pediatric Intensive Care Unit. en. Critical Care Nursing Clinics of North America 2005 Sep; 17:239–44. DOI: 10.1016/j.ccell.2005.04.005. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0899588505000481 [Accessed on: 2025 Jan 10]
4. AL-Samsam RH and Cullen P. Sleep and adverse environmental factors in sedated mechanically ventilated pediatric intensive care patients: en. Pediatric Critical Care Medicine 2005 Sep; 6:562–7. DOI: 10.1097/01.PCC.0000165561.40986.A6. Available from: http://journals.lww.com/00130478-200509000-00011 [Accessed on: 2025 Jan 10]
5. Weinhouse GL, Schwab RJ, Watson PL, Patil N, Vaccaro B, Pandharipande P, and Ely EW. Bench-to-bedside review: Delirium in ICU patients - importance of sleep deprivation. en. Critical Care 2009; 13:234. DOI: 10.1186/cc8131. Available from: http://ccforum.biomedcentral.com/articles/10.1186/cc8131 [Accessed on: 2025 Jan 10]
6. Vyazovskiy VV and Delogu A. NREM and REM Sleep: Complementary Roles in Recovery after Wakefulness. en. The Neuroscientist 2014 Jun; 20:203–19. DOI: 10.1177/1073858413518152. Available from: https://journals.sagepub.com/doi/10.1177/1073858413518152 [Accessed on: 2024 Sep 20]
7. Morse AM and Bender E. Sleep in Hospitalized Patients. en. Clocks & Sleep 2019 Feb; 1:151–65. DOI: 10.3390/clockssleep1010014. Available from: https://www.mdpi.com/2624-5175/1/1/14 [Accessed on: 2024 Sep 20]
8. Jafari B and Mohsenin V. Polysomnography. en. Clinics in Chest Medicine 2010 Jun; 31:287–97. DOI: 10.1016/j.ccm.2010.02.005. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0272523110000286 [Accessed on: 2024 Sep 20]
9. Berry RB, Brooks R, Gamaldo C, Harding SM, Lloyd RM, Quan SF, Troester MT, and Vaughn BV. AASM Scoring Manual Updates for 2017 (Version 2.4). en. Journal of Clinical Sleep Medicine 2017 May; 13:665–6. DOI: 10.5664/jcsm.6576. Available from: http://jcsm.aasm.org/doi/10.5664/jcsm.6576 [Accessed on: 2024 Sep 20]
10. Le Bon O. Relationships between REM and NREM in the NREM-REM sleep cycle: a review on competing concepts. en. Sleep Medicine 2020 Jun; 70:6–16. DOI: 10.1016/j.sleep.2020.02.004. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1389945720300757 [Accessed on: 2024 Sep 27]
11. Danker-Hopfe H, Kunz D, Gruber G, Klösch G, Lorenzo JL, Himanen SL, Kemp B, Penzel T, Röschke J, Dorn H, Schlögl A, Trenker E, and Dorffner G. Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. en. Journal of Sleep Research 2004 Mar; 13:63–9. DOI: 10.1046/j.1365-2869.2003.00375.x. Available from: https://onlinelibrary.wiley.com/doi/10.1046/j.1365-2869.2003.00375.x [Accessed on: 2024 Sep 27]
12. Younes M, Raneri J, and Hanly P. Staging Sleep in Polysomnograms: Analysis of Inter-Scorer Variability. en. Journal of Clinical Sleep Medicine 2016 Jun; 12:885–94. DOI: 10.5664/jcsm.5894. Available from: http://jcsm.aasm.org/doi/10.5664/jcsm.5894 [Accessed on: 2024 Sep 27]
13. Norman RG, Pal I, Stewart C, Walsleben JA, and Rapoport DM. Interobserver Agreement Among Sleep Scorers From Different Centers in a Large Dataset. en. Sleep 2000 Oct; 23:1–8. DOI: 10.1093/sleep/23.7.1e. Available from: https://academic.oup.com/sleep/article/23/7/1/2753224 [Accessed on: 2024 Sep 27]

14. Jager M, Tax D, Jonge R de, Kuiper JW, Twist E van, and Winder B van. Sleep Stage Classification Using Supervised Dynamic Machine Learning Models on EEG Data: A Scoping Review. Not published 2024 Oct

15. Mirmiran M, Maas YG, and Ariagno RL. Development of fetal and neonatal sleep and circadian rhythms. en. Sleep Medicine Reviews 2003 Aug; 7:321–34. DOI: 10.1053/smrv.2002.0243. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1087079202902431 [Accessed on: 2024 Oct 15]

16. Horne J. Why REM sleep? Clues beyond the laboratory in a more challenging world. en. Biological Psychology 2013 Feb; 92:152–68. DOI: 10.1016/j.biopsycho.2012.10.010. Available from: https://linkinghub.elsevier.com/retrieve/pii/S030105111200230X [Accessed on: 2024 Oct 15]

17. Lenard H. The Development of Sleep Spindles in the EEG During the First Two Years of Life. en. Neuropediatrics 1970 Feb; 1:264–76. DOI: 10.1055/s-0028-1091818. Available from: http://www.thieme-connect.de/DOI/DOI?10.1055/s-0028-1091818 [Accessed on: 2025 Feb 6]

18. Grigg-Damberger M, Gozal D, Marcus CL, Quan SF, Rosen CL, Chervin RD, Wise M, Picchietti DL, Sheldon SH, and Iber C. The Visual Scoring of Sleep and Arousal in Infants and Children. en. Journal of Clinical Sleep Medicine 2007 Mar; 03:201–40. DOI: 10.5664/jcsm.26819. Available from: http://jcsm.aasm.org/doi/10.5664/jcsm.26819 [Accessed on: 2025 Feb 6]

19. Matejcek M, Pokorny R, Ferber G, and Klee H. Effect of Morphine on the Electroencephalogram and Other Physiological and Behavioral Parameters. en. Neuropsychobiology 1988; 19:202–11. DOI: 10.1159/000118461. Available from: https://karger.com/NPS/article/doi/10.1159/000118461 [Accessed on: 2025 Feb 6]

20. Lancel M. Role of GABAA Receptors in the Regulation of Sleep: Initial Sleep Responses to Peripherally Administered Modulators and Agonists. en. Sleep 1999 Jan; 22:33–42. DOI: 10.1093/sleep/22.1.33. Available from: https://academic.oup.com/sleep/article-lookup/doi/10.1093/sleep/22.1.33 [Accessed on: 2025 Feb 6]

21. Veselis RA, Reinsel R, Marino P, Sommer S, and Carlon GC. The effects of midazolam on the EEG during sedation of critically ill patients. en. Anaesthesia 1993 Jun; 48:463–70. DOI: 10.1111/j.1365-2044.1993.tb07063.x. Available from: https://associationofanaesthetists-publications.onlinelibrary.wiley.com/doi/10.1111/j.1365-2044.1993.tb07063.x [Accessed on: 2025 Feb 6]

22. Nuwer MR, Hovda DA, Schrader LM, and Vespa PM. Routine and quantitative EEG in mild traumatic brain injury. en. Clinical Neurophysiology 2005 Sep; 116:2001–25. DOI: 10.1016/j.clinph.2005.05.008. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1388245705002130 [Accessed on: 2025 Feb 6]

23. Mantua J, Grillakis A, Mahfouz SH, Taylor MR, Brager AJ, Yarnell AM, Balkin TJ, Capaldi VF, and Simonelli G. A systematic review and meta-analysis of sleep architecture and chronic traumatic brain injury. en. Sleep Medicine Reviews 2018 Oct; 41:61–77. DOI: 10.1016/j.smrv.2018.01.004. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1087079217301752 [Accessed on: 2025 Feb 6]

24. Young GB. The EEG in Coma: en. Journal of Clinical Neurophysiology 2000 Sep; 17:473–85. DOI: 10.1097/00004691-200009000-00006. Available from: http://journals.lww.com/00004691-200009000-00006 [Accessed on: 2025 Feb 6]

25. Kaplan PW. The EEG in Metabolic Encephalopathy and Coma. Journal of Clinical Neurophysiology 2004 Sep; 21:307–18

26. Kemp B. https://www.physionet.org/content/sleep-edfx/1.0.0/

27. Van Twist E, Hiemstra FW, Cramer AB, Verbruggen SC, Tax DM, Joosten K, Louter M, Straver DC, De Hoog M, Kuiper JW, and De Jonge RC. An electroencephalography-based sleep index and supervised machine learning as a suitable tool for automated sleep classification in children. en. Journal of Clinical Sleep Medicine 2024 Mar; 20:389–97. DOI: 10.5664/jcsm.10880. Available from: http://jcsm.aasm.org/doi/10.5664/jcsm.10880 [Accessed on: 2025 Jan 10]

28. Cramer AB. Children in the pediatric intensive care unit experience limited REM sleep and frequent awakenings, and exhibit atypical electroencephalograms. Unpublished work. 2023.

29. Veldscholte K, Cramer AB, De Jonge RC, Rizopoulos D, Joosten KF, and Verbruggen SC. Intermittent feeding with an overnight fast versus 24-h feeding in critically ill neonates, infants, and children: An open-label, single-centre, randomised controlled trial. en. Clinical Nutrition 2023 Sep; 42:1569–80. DOI: 10.1016/j.clnu.2023.07.010. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0261561423002297 [Accessed on: 2025 Feb 3]

30. Supratak A, Dong H, Wu C, and Guo Y. DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG. IEEE Transactions on Neural Systems and Rehabilitation Engineering 2017 Nov; 25:1998–2008. DOI: 10.1109/TNSRE.2017.2721116. Available from: https://ieeexplore.ieee.org/document/7961240/ [Accessed on: 2024 Sep 30]

31. Hiemstra FW. Automated EEG-based sleep monitoring in critically ill children. PhD thesis. 2021 Sep

32. Sekkal RN, Bereksi-Reguig F, Ruiz-Fernandez D, Dib N, and Sekkal S. Automatic sleep stage classification: From classical machine learning methods to deep learning. en. Biomedical Signal Processing and Control 2022 Aug; 77:103751. DOI: 10.1016/j.bspc.2022.103751. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1746809422002737 [Accessed on: 2024 Sep 27]

33. Yazdi M, Samaee M, and Massicotte D. A Review on Automated Sleep Study. en. Annals of Biomedical Engineering 2024 Jun; 52:1463–91. DOI: 10.1007/s10439-024-03486-0. Available from: https://link.springer.com/10.1007/s10439-024-03486-0 [Accessed on: 2024 Sep 27]

34. Alsolai H, Qureshi S, Iqbal SMZ, Vanichayobon S, Henesey LE, Lindley C, and Karrila S. A Systematic Review of Literature on Automated Sleep Scoring. IEEE Access 2022; 10:79419–43. DOI: 10.1109/ACCESS.2022.3194145. Available from: https://ieeexplore.ieee.org/document/9841539/ [Accessed on: 2024 Sep 27]

35. Fiorillo L, Wand M, Marino I, Favaro P, and Faraci FD. Temporal dependency in automatic sleep scoring via deep learning based architectures: An empirical study. Annu Int Conf IEEE Eng Med Biol Soc 2020; 2020:3509–12. DOI: 10.1109/embc44109.2020.9176356. Available from: http://dx.doi.org/10.1109/embc44109.2020.9176356%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=33018760%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med18&DO=10.1109%2fEMBC44109.2020.9176356 [Accessed on: 9 Jan 1]

36. Chambon S, Galtier MN, Arnal PJ, Wainrib G, and Gramfort A. A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series. IEEE Trans Neural Syst Rehabil Eng 2004; 26:758–69. DOI: 10.1109/tnsre.2018.2813138. Available from: http://dx.doi.org/10.1109/tnsre.2018.2813138%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=29641380%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med15&DO=10.1109%2fTNSRE.2018.2813138

37. Pan ST, Kuo CE, Zeng JH, and Liang SF. A transition-constrained discrete hidden Markov model for automatic sleep staging. Biomed. eng. online 2012; 11:52. DOI: 10.1186/1475-925x-11-52. Available from: http://dx.doi.org/10.1186/1475-925x-11-52%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=22908930%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med9&DO=10.1186%2f1475-925X-11-52 [Accessed on: 8 Jan 1]

38. Ghimatgar H, Kazemi K, Helfroush MS, and Aarabi A. An automatic single-channel EEG-based sleep stage scoring method based on hidden Markov Model. J Neurosci Methods 2019; 324:108320. DOI: 10.1016/j.jneumeth.2019.108320. Available from: http://dx.doi.org/10.1016/j.jneumeth.2019.108320%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=31228517%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med16&DO=10.1016%2fj.jneumeth.2019.108320 [Accessed on: 9 Jan 1]

39. Wang X and Zhu Y. SleepGCN: A transition rule learning model based on Graph Convolutional Network for sleep staging. Comput Methods Programs Biomed 2024; 257:108405. DOI: 10.1016/j.cmpb.2024.108405. Available from: http://dx.doi.org/10.1016/j.cmpb.2024.108405%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=inleurlib_fft&cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=39243591%20https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=medp&DO=10.1016%2fj.cmpb.2024.108405 [Accessed on: 9 Jan 1]

40. Hipp JF and Siegel M. Dissociating neuronal gamma-band activity from cranial and ocular muscle activity in EEG. Frontiers in Human Neuroscience 2013; 7. DOI: 10.3389/fnhum.2013.00338. Available from: http://journal.frontiersin.org/article/10.3389/fnhum.2013.00338/abstract [Accessed on: 2025 Mar 25]

41. Korkalainen H, Leppanen T, Aakko J, Nikkonen S, Kainulainen S, Leino A, Duce B, Afara IO, Myllymaa S, and Toyras J. Accurate Deep Learning-Based Sleep Staging in a Clinical Population with Suspected Obstructive Sleep Apnea. IEEE Journal of Biomedical and Health Informatics 2019 :1–1. DOI: 10.1109/JBHI.2019.2951346. Available from: https://ieeexplore.ieee.org/document/8936942/ [Accessed on: 2025 Mar 25]

42. Ambrogio C, Koebnick J, Quan SF, Ranieri VM, and Parthasarathy S. Assessment of Sleep in Ventilator-Supported Critically Ill Patients. en. Sleep 2008 Nov; 31:1559–68. DOI: 10.1093/sleep/31.11.1559. Available from: https://academic.oup.com/sleep/article-lookup/doi/10.1093/sleep/31.11.1559 [Accessed on: 2025 Mar 30]

43. Elliott R, McKinley S, Cistulli P, and Fien M. Characterisation of sleep in intensive care using 24-hour polysomnography: anobservational study. en. Critical Care 2013 Mar; 17:R46. DOI: 10.1186/cc12565. Available from: https://ccforum.biomedcentral.com/articles/10.1186/cc12565 [Accessed on: 2025 Mar 30]

44. Sun H, Ganglberger W, Panneerselvam E, Leone MJ, Quadri SA, Goparaju B, Tesh RA, Akeju O, Thomas RJ, and Westover MB. Sleep staging from electrocardiography and respiration with deep learning. en. Sleep 2020 Jul; 43:zsz306. DOI: 10.1093/sleep/zsz306. Available from: https://academic.oup.com/sleep/article/doi/10.1093/sleep/zsz306/5682785 [Accessed on: 2025 Mar 30]

45. Urtnasan E, Park JU, Joo EY, and Lee KJ. Deep Convolutional Recurrent Model for Automatic Scoring Sleep Stages Based on Single-Lead ECG Signal. en. Diagnostics 2022 May; 12:1235. DOI: 10.3390/diagnostics12051235. Available from: https://www.mdpi.com/2075-4418/12/5/1235 [Accessed on: 2025 Mar 30]

46. Van Twist E, Meester AM, Cramer ABG, De Hoog M, Schouten AC, Verbruggen SCAT, Joosten KFM, Louter M, Straver DCG, Tax DMJ, De Jonge RCJ, and Kuiper JW. Supervised machine learning on electrocardiography features to classify sleep in noncritically ill children. en. Journal of Clinical Sleep Medicine 2025 Feb; 21:261–8. DOI: 10.5664/jcsm.11358. Available from: http://jcsm.aasm.org/doi/10.5664/jcsm.11358 [Accessed on: 2025 Apr 17]

S. Supplementary Materials

S.1 Detailed Characteristics of the Patients of the PICU Dataset

Table 13: Detailed patient characteristics of the patient from the PICU dataset.

Patient	PICU Day	Age (days)	Age Group	Gender	Diagnosis Group	Neurological Condition	PELOD-2 Score	PIM3 Score	Medication
PICU001	2	3	0-2 months	f	Abdominal Surgery	-	3	4.89	-
PICU002	2	3	0-2 months	f	Respiratory	-	9	1.76	Midazolam, Opioids
PICU003	8	9	0-2 months	f	Respiratory	Asphyxia, hyperechogenic periventricular white matter abnormalities	7	0.52	Midazolam, Opioids, Esketamine
PICU004	2	6	0-2 months	f	Metabolic	Metabolic encephalopathy	6	1.65	Opioids, Esketamine
PICU005	2	8	0-2 months	f	Cardiac	-	8	1.97	Opioids
PICU006	13	24	0-2 months	f	Cardiac	-	5	1.87	Midazolam
PICU007	8	20	0-2 months	f	Cardiac	-	2	-3.65	-
PICU008	8	26	0-2 months	m	Cardiac	-	2	-2.53	Midazolam, Opioids
PICU009	13	31	0-2 months	m	Neurological	Floppy Infant Syndrome (etiology unclear)	7	3.26	Midazolam
PICU010	2	39	0-2 months	m	Respiratory	-	5	4.7	-
PICU011	2	51	0-2 months	m	Respiratory	-	7	4.9	Midazolam, Opioids
PICU012	3	62	2-6 months	f	Cardiac	-	9	4	Midazolam, Esketamine
PICU013	6	72	2-6 months	m	Infectious	-	11	4.28	Midazolam, Opioids, Esketamine
PICU014	2	81	2-6 months	m	Respiratory	Encephalopathy, atrophy of frontal and temporal lobes and corpus callosum, pyridoxine-dependent epilepsy	7	2.92	-
PICU015	7	109	2-6 months	f	Cardiac	-	9	3.61	Midazolam, Opioids, Esketamine
PICU016	7	117	2-6 months	f	Cardiac	Encephalopathy, accumulation of sedatives	14	4.06	Midazolam, Opioids
PICU017	14	131	2-6 months	m	Cardiac	-	9	4.56	Midazolam, Opioids
PICU018	4	132	2-6 months	f	Cardiac	-	4	4.45	Midazolam
PICU019	7	222	6-12 months	f	Cardiac	-	21	1.41	Midazolam
PICU020	1	231	6-12 months	m	Cardiac	-	3	4.98	Opioids

PICU = paediatric intensive care unit, PELOD-2 = paediatric logistic organ dysfunction 2, PIM3 = paediatric index of mortality 3, f = female, m = male.

Table 14: Detailed patient characteristics of the patient from the PICU dataset.

Patient	PICU Day	Age (days)	Age Group	Gender	Diagnosis Group	Neurological Condition	PELOD-2 Score	PIM3 Score	Medication
PICU021	1	402	1-3 years	m	Respiratory	-	7	2.68	Midazolam, Opioids
PICU022	7	623	1-3 years	m	Respiratory	Prader-Willi Syndrome	7	4.96	Midazolam, Opioids
PICU023	3	2197	5-9 years	f	Oncological	-	7	4.70	Midazolam, Opioids, Esketamine
PICU024	3	3675	9-13 years	f	Infectious	Optic neuritis with multiple white matter abnormalities	7	3.91	Midazolam, Opioids
PICU025	3	4767	13-18 years	m	Neurological	Encephalopathy, post-resuscitation and intra parenchymal haemorrhage in the cerebellum	8	4.03	Opioids
PICU026	12	5288	13-18 years	m	Infectious	Encephalopathy, post-resuscitation and small right parietal punctate haemorrhage	9	3.57	Opioids, Esketamine
PICU027	2	6091	13-18 years	f	Neurological	Myasthenia Gravis	5	1.73	Opioids
PICU028	2	6311	13-18 years	m	Infectious	Encephalopathy, pontine ischaemia	9	1.33	Midazolam, Opioids

PICU = paediatric intensive care unit, PELOD-2 = paediatric logistic organ dysfunction 2, PIM3 = paediatric index of mortality 3, f = female, m = male.

S.2 Explanation of a Convolutional Neural Network

A CNN is a type of neural network designed to process structured data by using layers like convolutional layers for feature extraction, pooling layers for dimensionality reduction, and fully connected layers for decision-making. Figure 18 gives an example of the flow of data through these layers, showing how CNNs transform raw input into predictions.

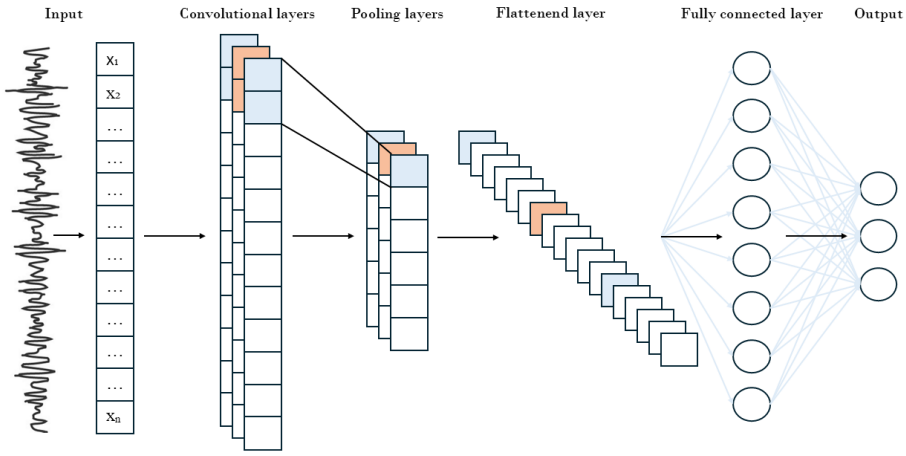


Figure 18: Architecture of a Convolutional Neural Network. It visualises how data flows through different layer types. Blocks represent feature maps or vectors, and circles represent neurons or a final probability.

Convolutional Layer

A convolutional layer applies a set of filters to the input data to extract meaningful features. Each filter moves across the input data with a specified stride size, performing element-wise multiplication between the filter values and the corresponding input values. The results are summed at each position, producing an output feature map that emphasises detected patterns.

The convolution operation is defined as:

$$y(i,k) = \sum_{m=0}^{F} x(i+m) \cdot w(m,k)$$

where:

- $y(i,k)$: The output at position i for the k-th filter
- $x(i+m)$: The input signal at position $i+m$
- $w(m,k)$: The filter of size F at position m for the k-th filter

Filter size: Determines the dimensions of the filters. Smaller filters capture fine-grained details, while larger filters detect broader patterns.

Number of filters: Specifies how many feature maps are generated. Each filter extracts different types of features from the input.

Stride size: Defines the step size of the filter movement. Larger strides reduce the resolution of the output.

Batch Normalization

Batch normalisation stabilises and accelerates training by normalising the input batch's mean and variance. The normalised input is scaled and shifted using learnable parameters:

$$y = \gamma \cdot \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta$$

where:

- x: The input
- μ: The mean of the input batch
- σ^2: The variance of the input batch
- ϵ: A small constant for numerical stability
- γ and β : Learnable scaling parameters
- y: The output after normalization.

Rectified Linear Unit (ReLU) Function

The ReLU function introduces non-linearity by allowing only positive values to pass through, setting negative values to 0. Batch normalisation and ReLU functions are performed after after/within a convolutional layer. It is defined as:

$$y = \max(0, x)$$

where:

- x: The input,
- y: The output, which equals x if $x > 0$, and 0 otherwise.

Pooling Layer

Pooling layers reduce the spatial dimensions of feature maps while retaining key information. Max pooling, the most common type, selects the maximum value in each pooling region:

$$y(i) = \max_{m \in R} \{x(i + m)\}$$

where:

- $y(i)$: The output at position i
- $x(i + m)$: The input signal at position $i + m$
- R: The pooling region

Pooling size: Defines the dimensions of the pooling region. Larger pooling sizes reduce dimensionality more aggressively but may lose fine details.

Stride size: Specifies how far the pooling region moves at each step, similar to the stride in convolutional layers.

Concatenate Layer

A concatenate layer combines multiple feature maps along a specified axis, creating a larger feature map. For example, given two feature maps v_1 and v_2, the concatenated output (y) is:

$$y = [v_1, v_2]$$

Flatten Layer

After previous layers, the output consists of multiple feature maps. The flatten layer converts these multi-dimensional arrays into a one-dimensional vector, enabling the data to be passed into fully connected layers for classification.

Dense Layer/ Fully Connected Layer

A dense or fully connected layer transforms input features into probabilities using a linear transformation followed by an activation function. It combines all previously learned features and makes class probabilities for the final prediction. The dense layer is defined as:

$$y = \sigma(Wx + b)$$

where:

- y: The output probabilities for each class
- x: The input feature map
- W: The weight matrix learned during training
- b: The bias vector
- σ: The activation function

Softmax Function

The softmax function converts raw scores (logits) into probabilities that sum to 1. It can be used as activation function in a dense layer:

$$y_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

where:

- y_i: The output probability for the i-th class
- z_i: The input for the i-th class
- $\exp(z_i)$: The exponential of z_i
- $\sum_j \exp(z_j)$: The sum of exponentials for all inputs

S.3 Explanation of a Long Short-Term Memory Model

LSTMs are designed to model sequential data by capturing short- and long-term dependencies. When used in combination with a CNN, the LSTM takes the output features extracted by the CNN as input and learns temporal patterns, enabling it to make predictions based on the sequential nature of the data. Figure 19 gives an example of the flow of data through an LSTM. On the right, a node of the LSTM-layer is highlighted to illustrate how information is passed on over time.

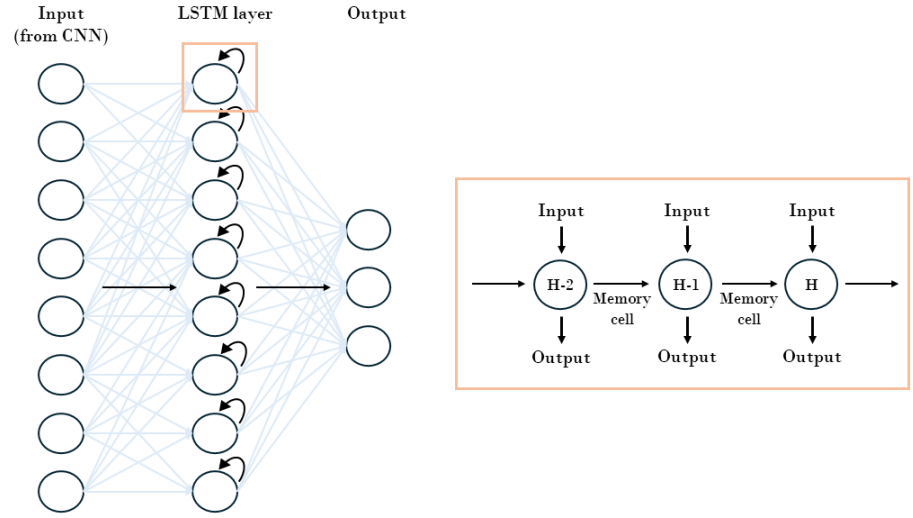


Figure 19: Architecture of a Long Short-Term Memory Model. It visualises how temporal information is passed on over time. Circles represent neurons or a final probability.

LSTM Layer

An LSTM layer processes input data across time steps and learns to retain important information. It consists of memory cells with gates that regulate the flow of information, thereby limiting the vanishing gradient problem, where long-term information is lost over time, seen in other recurrent neural networks.

In the LSTM layers the following parameters are used:

- **Input** (x_t): The data at the current time step.
- **Hidden State** (h_t): The output of the LSTM at the current time step.
- **Cell State/Memory Cell** (c_t): The internal memory of the LSTM that stores long-term information across time steps. The cell state is updated using the forget, input, and candidate gates.
- **Gates** (f_t, i_t, o_t): The forget, input, and candidate gates that regulate the flow of information
- **Weight Matrices** (W): The learned parameters that control how the input and hidden states are transformed at each gate.
- **Biases** (b): Bias terms added to each gate's computation to allow the network to shift activation thresholds.
- **Activation Functions** (σ and tanh): Non-linear functions used in the LSTM

The computations for a given time step t are done as follows:

Forget Gate: Controls how much of the previous cell state c_{t-1} should be retained. Values closer to 1 allow information to pass through, while values closer to 0 forget it.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

Input Gate: Decides how much new information from the current input x_t is added to the memory.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

Candidate Memory: Creates a candidate update for the memory cell state.

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

Cell State Update: Combines the retained memory ($f_t \odot c_{t-1}$) and the new information ($i_t \odot \tilde{c}_t$).

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

Output Gate: Determines how much of the updated cell state will contribute to the output.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

Hidden State: Produces the output of the LSTM cell, which is also passed to the next time step.

$$h_t = o_t \odot \tanh(c_t)$$

Output dimensionality: This refers to the size of hidden state vector (h_t) at each time step. This is a configurable parameter that determines the number of features learned and produced by the LSTM at each step in the sequence. A larger dimensionality allows the model to learn and represent more complex patterns but increases computational costs.

Element-Wise Addition

Element-wise addition combines two vectors or matrices of the same dimensions by adding their corresponding elements. This operation ensures that the outputs of the CNN through the fully connected layer and LSTM layers are combined.

The operation is defined as:

$$z(i) = y_1(i) + y_2(i)$$

where:

- $y_1(i)$: The i-th element of the first vector,
- $y_2(i)$: The i-th element of the second vector,
- $z(i)$: The resulting i-th element after addition.

Other Layers and Parameters

Dense layers, batch normalisation, and a ReLu function are explained in Appendix S.2.

S.4 Distribution of Principal Components of CNN-derived Features

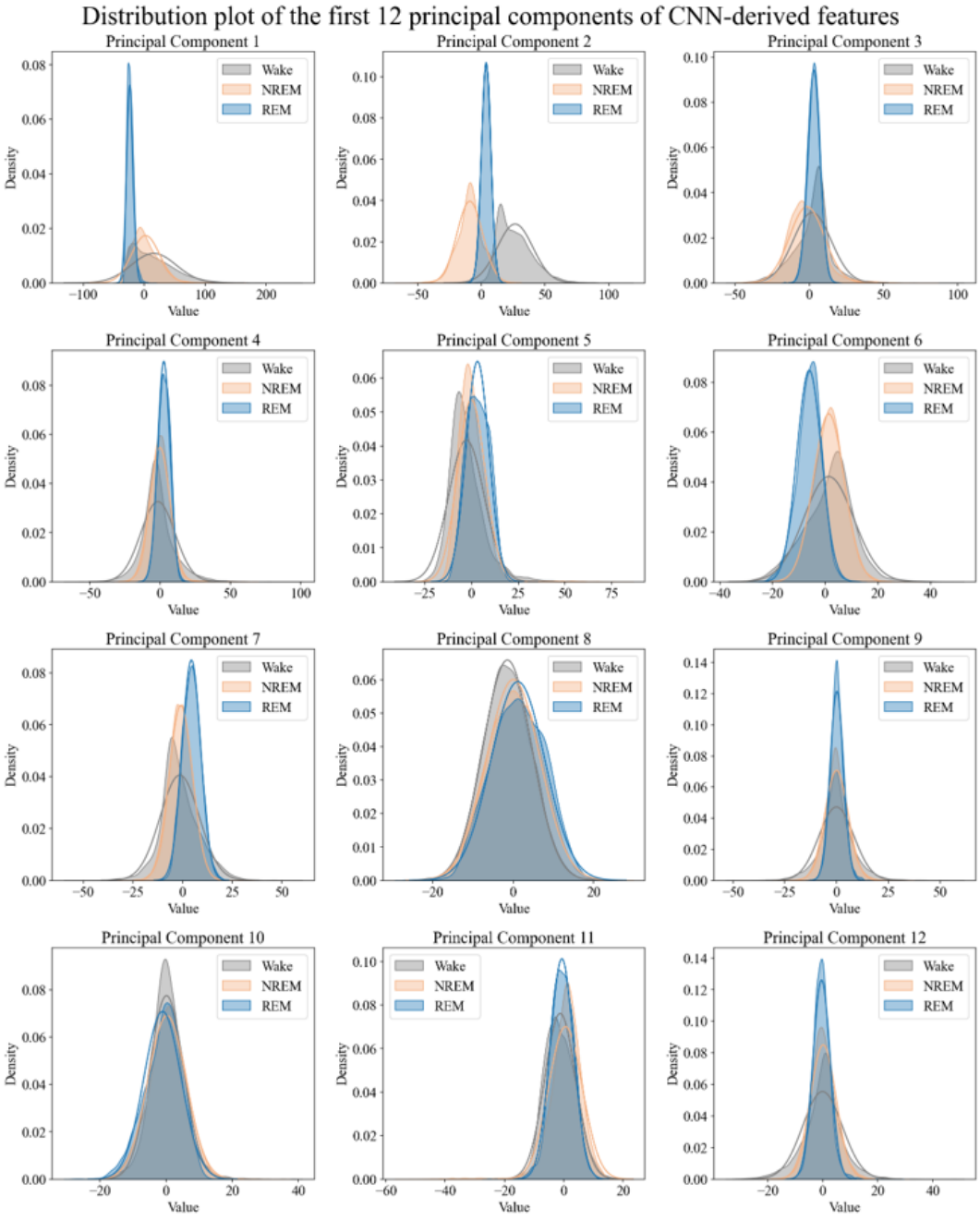


Figure 20: Distribution plots of the first 12 principal components of CNN-derived features from the Sleep-EDF dataset, stratified by sleep stage (Wake, NREM, REM). For each principal component, the coloured areas represent the actual distributions of the data, while the overlaid lines indicate fitted Gaussian distributions per sleep stage. Both reflect probability density functions. These plots illustrate whether features are normally distributed and the degree of overlap and separability between stages in the reduced feature space.

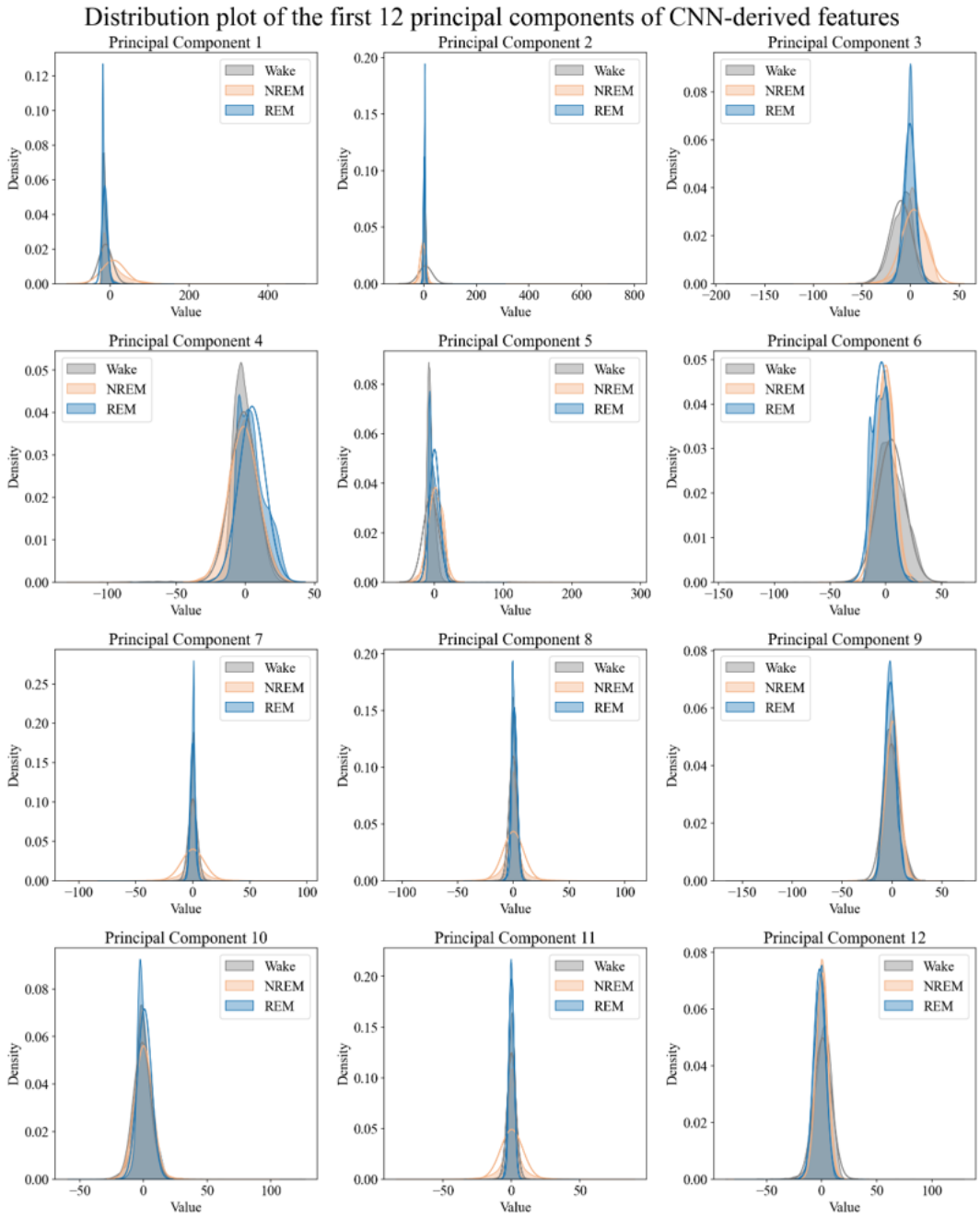


Figure 21: Distribution plots of the first 12 principal components of CNN-derived features from the PSG dataset, stratified by sleep stage (Wake, NREM, REM). For each principal component, the coloured areas represent the actual distributions of the data, while the overlaid lines indicate fitted Gaussian distributions per sleep stage. Both reflect probability density functions. These plots illustrate whether features are normally distributed and the degree of overlap and separability between stages in the reduced feature space.

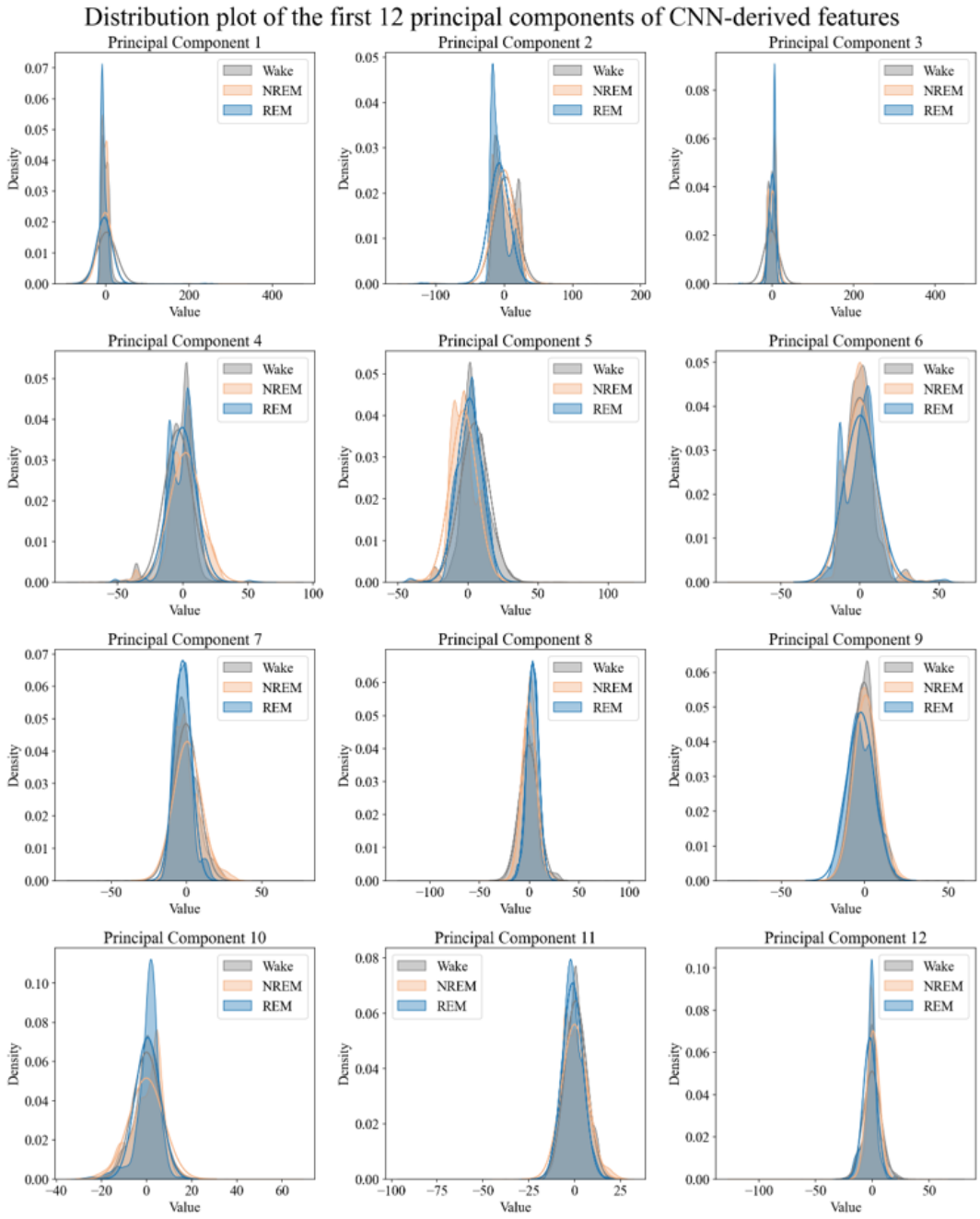


Figure 22: Distribution plots of the first 12 principal components of CNN-derived features from the PICU dataset, stratified by sleep stage (Wake, NREM, REM). For each principal component, the coloured areas represent the actual distributions of the data, while the overlaid lines indicate fitted Gaussian distributions per sleep stage. Both reflect probability density functions. These plots illustrate whether features are normally distributed and the degree of overlap and separability between stages in the reduced feature space.

S.5 Explanation of a Hidden Markov Model

An HMM is a probabilistic model used to represent sequential data by modeling transitions between hidden states. Figure 18 illustrates the structure of an HMM, containing the following components:

- Hidden states (orange): The sleep stages in which the model classifies epochs
- Observed states (blue): The data that can be measured, in this case, EEG features
- Transition probabilities (black lines): The probability of moving from one hidden state to another
- Emission probabilities (grey lines): The probability of observing an observed state given a hidden state
- Initial probabilities (not shown): The probability that a sequence start in a given state

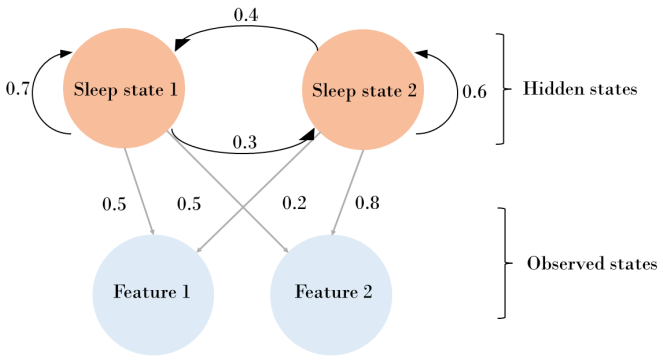


Figure 23: Architecture of the Hidden Markov Model. It consists of hidden states representing different sleep stages and observed states representing features derived from the CNN. Transition probabilities between hidden states (black arrows) and emission probabilities from hidden to observed states (grey arrows) are modelled.

Transition Probabilities

The state transition probability matrix A defines the probability of transitioning from state s_i to s_j:

$$a_{ij} = P(S_{t+1} = s_j | S_t = s_i)$$

where:

- a_{ij}: The probability of transitioning from state s_i to state s_j,
- S_t: The state at time t,

The sum of transition probabilities from each state must equal 1.

Emission Probabilities

Each hidden state generates an observable output according to the emission probability distribution, modelled using Gaussian distributions. The emission probability matrix B is defined as:

$$b_j(o_t) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_j|}} \exp\left(-\frac{1}{2}(o_t - \mu_j)^\top \Sigma_j^{-1}(o_t - \mu_j)\right)$$

where:

- $b_j(o_t)$: The probability density of observing o_t given the system is in state s_j.
- o_t: The observed feature vector at time t,
- μ_j: The mean vector of the Gaussian distribution associated with state s_j,
- Σ_j: The covariance matrix for state s_j,

Initial Probabilities

The initial probability π defines the probability of starting in each state, defined as:

$$\pi_i = P(S_1 = s_i)$$

where S_1 is the initial state. The sum of the initial probabilities must equal 1.

The Viterbi Algorithm

The Viterbi algorithm finds the most probable sequence of hidden states given a sequence of observations. It is defined as:

$$\delta_t(j) = \max_i \left[\delta_{t-1}(i) a_{ij}\right] b_j(O_t)$$

where:

- $\delta_t(j)$: The highest probability of reaching state j at time t,
- a_{ij}: The transition probability,
- $b_j(O_t)$: The emission probability.

The algorithm traces back the optimal state sequence using a backtracking step.

Training: Them Baum-Welch Algorithm

The Baum-Welch algorithm is an expectation-maximization (EM) technique used to train an HMM by estimating A, B, and π from observed sequences.

Expectation Step (E-Step): Computes the probability of transitioning between states based on observed sequences.

Maximization Step (M-Step): Updates the model parameters based on the probabilities computed in the E-step:

$$a_{ij}^{new} = \frac{\sum_{t=1}^{T-1} P(S_t = s_i, S_{t+1} = s_j | O, \lambda)}{\sum_{t=1}^{T-1} P(S_t = s_i | O, \lambda)}$$

$$b_j(o)^{new} = \frac{\sum_{t=1}^{T} P(S_t = s_j | O, \lambda) \cdot I(O_t = o)}{\sum_{t=1}^{T} P(S_t = s_j | O, \lambda)}$$

where $I(O_t = o)$ is an indicator function that is 1 when $O_t = o$ and 0 otherwise. The algorithm iterates until convergence.

S.6 Calculation of Performance Metrics

Accuracy (Acc): Represents the proportion of correctly classified instances:

$$ACC = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}.$$

Per-Class F1 Score: Reflects a balanced measure of precision and recall per class:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

where:

- $\text{Precision} = \frac{\text{True Positives}}{\text{True Positives+False Positives}}$,
- $\text{Recall} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$.

Macro-Averaged F1 Score (MF1): Calculates the average F1 score across all classes:

$$MF1 = \frac{1}{C} \sum_{i=1}^{C} F1_i$$

where:

- C: The total number of classes,
- $F1_i$: The F1 score for the i-th class.

Cohen's Kappa (Kappa): Reflects agreement between predicted and actual classifications while accounting for chance-level agreement:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where:

- P_o: Observed agreement (proportion of correctly classified instances),
- P_e: Expected agreement due to chance.

Area Under the Receiver-Operating Characteristic Curve (AUC-ROC): Measures the ability of the model to distinguish between classes and is calculated as the area under the ROC curve:

$$AUC = \int_0^1 TPR(FPR)\, dFPR$$

where:

- TPR: True positive rate (Recall),
- FPR: False positive rate.

Adjusted Rand Index (ARI): Measures the similarity between two predicted cluster assignments while adjusting for chance:

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \Big/ \binom{n}{2}\right]}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \Big/ \binom{n}{2}\right]}$$

where:

- n: Total number of samples,
- n_{ij}: Number of samples assigned to cluster i in the first prediction and cluster j in the second prediction,
- a_i: Number of samples in cluster i in the first prediction,
- b_j: Number of samples in cluster j in the second prediction.

ARI ranges from -1 (no agreement) to 1 (perfect agreement), with 0 representing agreement expected by chance.

Log-Likelihood: Measures how well a probabilistic model explains the observed data:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \log P(x_i \mid \theta)$$

where:

- n: Number of data points,
- $P(x_i \mid \theta)$: Probability of observing x_i given model parameters θ.

A higher log-likelihood value indicates a better fit of the model to the data. Since probabilities always take values between 0 and 1, their logarithm is negative. This means that each term in the sum contributes a negative value, making the log-likelihood typically negative. Moreover, because the log-likelihood sums over all data points, the total value decreases (becomes more negative) as the dataset grows.

In clustering, log-likelihood is often used to compare different probabilistic models based on how well they explain the observed data. Models with a less negative log-likelihood are preferred, as they indicate a better fit.

S.7 Influence of Input Channels on the Performance of the CNN

Table 15: Performance metrics of the convolutional neural network for three-state classification for the PSG dataset utilising different input channels. Results were obtained by 5-fold cross-validation.

Mean (SD) metric in %	EEG	EEG & EOG	EEG, EOG & EMG
Accuracy	71.3 (1.7)	72.0 (3.4)	71.1 (2.4)
Macro F1	69.2 (1.9)	68.4 (3.4)	67.9 (3.7)

SD = standard deviation, EEG = electroencefalography, EOG = electrooculography, EMG = electromyography.

S.8 Influence of Sequence Length on the Performance of the LSTM

Table 16: Performance metrics of the long short-term memory model for three-state classification on the PSG dataset utilising different sequence lengths as input. Results were obtained by 5-fold cross-validation. An epoch corresponds to 30 seconds of electroencephalography data.

Mean (SD) metric in %	5 epochs	10 epochs	20 epochs	60 epochs	120 epochs	240 epochs
Accuracy	75.9 (0.8)	76.3 (0.9)	77.4 (0.7)	76.8 (0.6)	78.0 (0.6)	76.1 (0.7)
Macro F1	72.1 (1.1)	72.6 (1.0)	73.5 (1.0)	73.0 (1.2)	74.0 (0.9)	72.5 (1.0)

SD = standard deviation

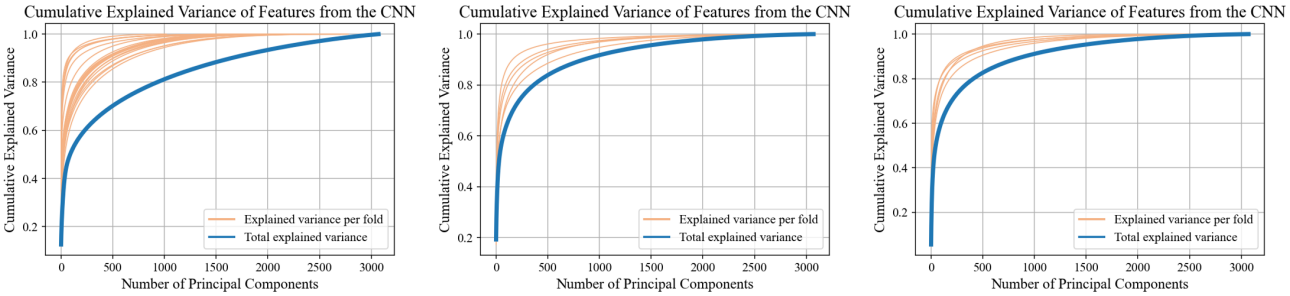
S.9 Influence of the Number of Principal Components on the Performance of the HMM

Table 17: Performance metrics of the hidden Markov model for three-state classification on the PSG dataset utilising a different number of principal components. Results were obtained by 5-fold cross-validation.

Mean (SD) metric in %	10 PCs	30 PCs	50 PCs	100 PCs	200 PCs	500 PCs
Accuracy	80.6 (7.4)	80.6 (6.6)	80.6 (6.4)	79.9 (6.6)	79.6 (6.8)	79.3 (7.0)
Macro F1	73.4 (7.7)	73.7 (7.4)	73.7 (6.9)	72.6 (7.3)	72.2 (7.5)	70.3 (8.9)

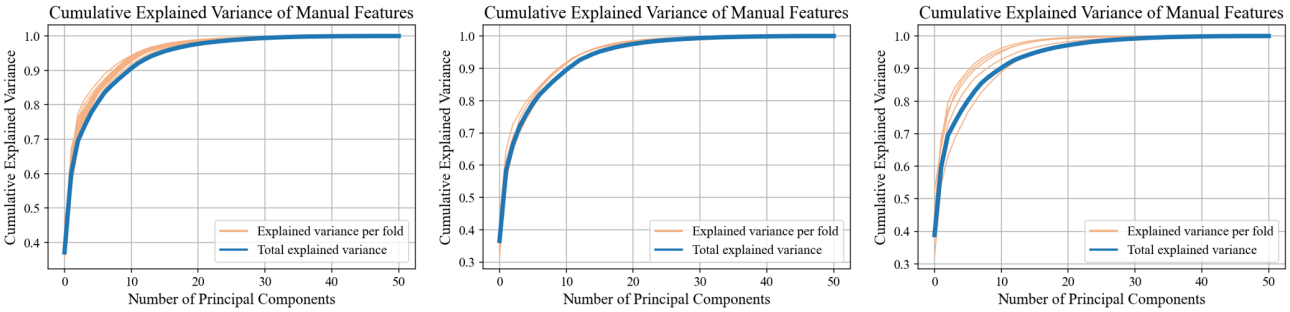
SD = standard deviation, PC = principal component

S.10 Explained Variance of Principal Components



(a) Sleep-EDF dataset **(b)** PSG dataset **(c)** PICU dataset

Figure 24: The explained variance by the principal components of the features derived from the convolutional neural network for the Sleep-EDF, PSG and PICU dataset. The results were obtained by a principal component analysis on each whole dataset (blue line) and for each fold of the 20- or 5-fold cross-validation for the SLeep-EDF and PSG/PICU dataset, respectively (orange lines).



(a) Sleep-EDF dataset **(b)** PSG dataset **(c)** PICU dataset

Figure 25: The explained variance by the principal components of the manually selected features for the Sleep-EDF, PSG and PICU dataset. The results were obtained by a principal component analysis on each whole dataset (blue line) and for each fold of the 20- or 5-fold cross-validation for the Sleep-EDF and PSG/PICU dataset, respectively (orange lines).

S.11 Confusion Matrices for Three- and Five-State Classification per Dataset

Table 18: Confusion matrix of the convolutional neural network on the Sleep-EDF dataset for three-state classification. The results are obtained by 20-fold cross-validation. The number of labels represents the total across all folds. F1-scores for each sleep stage are reported as the mean (SD) across folds.

Actual Labels	Predicted Labels			Mean F1 in % (SD)
	Wake	NREM	REM	
Wake	6509	521	383	85.0 (9.4)
NREM	1044	23489	1657	91.8 (3.6)
REM	270	1001	6426	79.4 (8.2)

SD = standard deviation, NREM = non rapid eye movement, REM = rapid eye movement

Table 19: Confusion matrix of the convolutional neural network on the Sleep-EDF dataset for five-state classification. The results are obtained by 20-fold cross-validation. The number of labels represents the total across all folds. F1-scores for each sleep stage are reported as the mean (SD) across folds.

Actual Labels	Predicted Labels					Mean F1 in % (SD)
	Wake	N1	N2	N3	REM	
Wake	5896	839	119	120	439	82.8 (10.9)
N1	299	1390	426	17	629	38.8 (11.4)
N2	348	893	14555	955	975	85.2 (5.5)
N3	23	14	364	5287	15	87.0 (6.3)
REM	192	950	514	22	6019	76.1 (8.7)

SD = standard deviation, REM = rapid eye movement, N1-N3 = non rapid eye movement stage 1-3.

Table 20: Confusion matrix of the convolutional neural network on the PSG dataset (9-18 years) for three-state classification. The results are obtained by 5-fold cross-validation. The number of labels represents the total across all folds. F1-scores for each sleep stage are reported as the mean (SD) across folds.

Actual Labels	Predicted Labels			Mean F1 in % (SD)
	Wake	NREM	REM	
Wake	5908	636	893	76.9 (4.8)
NREM	1602	16425	2162	86.9 (2.5)
REM	383	537	3254	63.1 (7.4)

SD = standard deviation, NREM = non rapid eye movement, REM = rapid eye movement

Table 21: Confusion matrix of the convolutional neural network on the PSG dataset (9-18 years) for five-state classification. The results are obtained by 50-fold cross-validation.The number of labels represents the total across all folds. F1-scores for each sleep stage are reported as the mean (SD) across folds.

Actual Labels	Predicted Labels					Mean F1 in % (SD)
	Wake	N1	N2	N3	REM	
Wake	4964	1000	238	142	1093	73.3 (6.5)
N1	633	1041	240	71	1247	32.1 (4.5)
N2	289	648	4683	1942	1159	63.9 (5.9)
N3	36	95	552	7349	177	82.6 (2.0)
REM	123	519	100	82	3350	60.3 (7.2)

SD = standard deviation, REM = rapid eye movement, N1-N3 = non rapid eye movement stage 1-3.

Table 22: Confusion matrix of the convolutional neural network on the PSG dataset for three-state classification. The results are obtained by 5-fold cross-validation. The number of labels represents the total across all folds. F1-scores for each sleep stage are reported as the mean (SD) across folds.

Actual Labels	Predicted Labels			Mean F1 in % (SD)
	Wake	NREM	REM	
Wake	16479	2410	10296	65.6 (6.2)
NREM	2888	64378	20821	81.2 (2.2)
REM	1223	3462	24743	58.3 (2.7)

SD = standard deviation, NREM = non rapid eye movement, REM = rapid eye movement

Table 23: Confusion matrix of the convolutional neural network on the PICU dataset for three-state classification. The results are obtained by 5-fold cross-validation. The number of labels represents the total across all folds. F1-scores for each sleep stage are reported as the mean (SD) across folds.

Actual Labels	Predicted Labels			Mean F1 in % (SD)
	Wake	NREM	REM	
Wake	12168	7620	5536	44.0 (19.4)
NREM	6074	28349	7483	69.7 (8.4)
REM	1408	2069	3593	24.0 (15.5)

SD = standard deviation, NREM = non rapid eye movement, REM = rapid eye movement

S.12 Distribution of Sleep Stages per Patient in the PICU Dataset

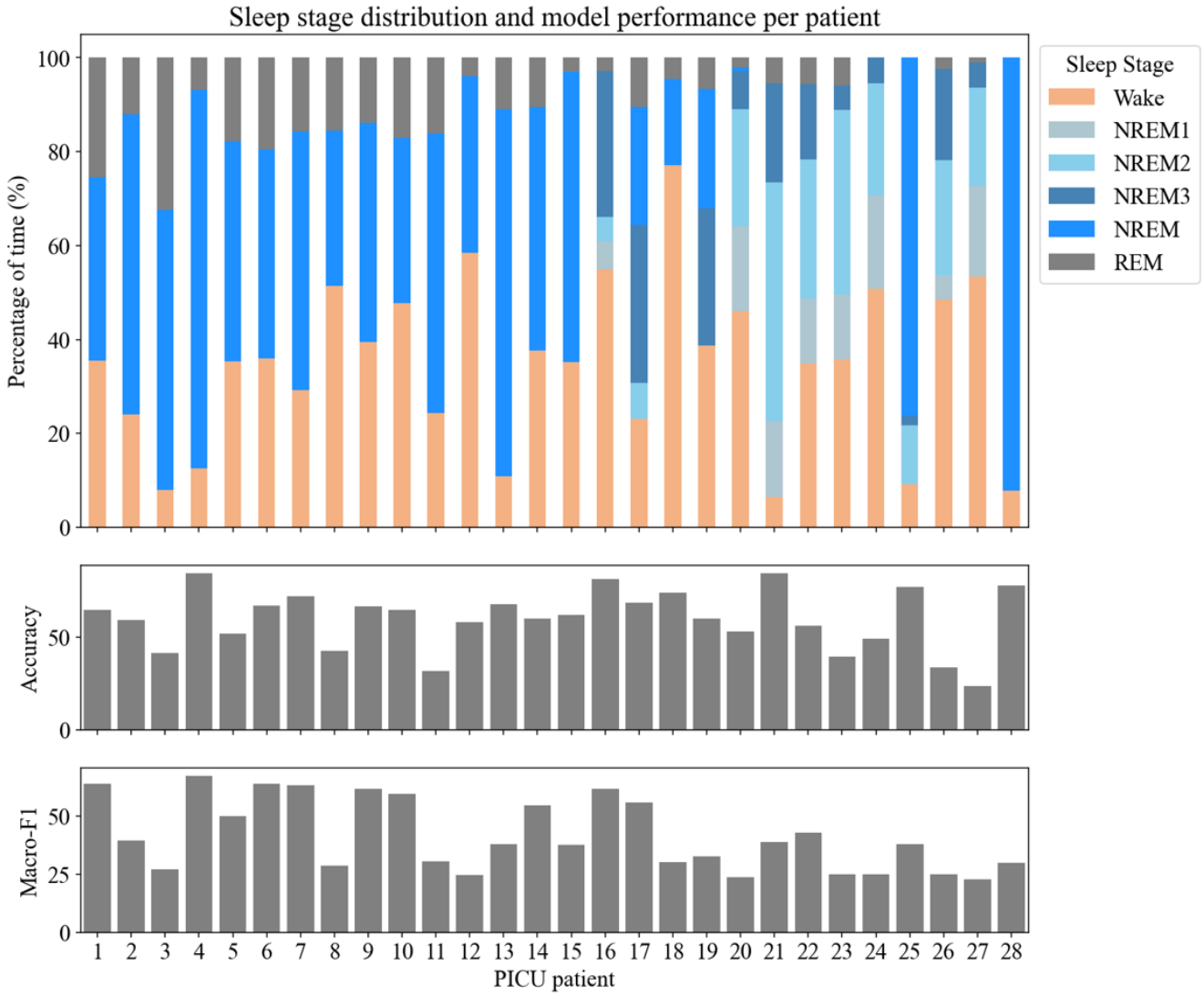


Figure 26: Distribution of manually labelled sleep stages for individual patients of the PICU Dataset. The first figure visualises the percentage of epochs that are classified within each individual patient. The bottom figures provide the accuracy and macro-F1 score achieved for this patient, obtained by the convolutional neural network through 5-fold cross-validation. NREM = non rapid eye movement, REM = rapid eye movement.

S.13 Performance of the CNN per Patient of the PICU Dataset with Varying Input

Table 24: Comparison between patient characteristics and three-state performance of the convolutional neural network on patient level Results were obtained by 5-fold cross-validation. Results are provided only by utilising EEG as an input channel or by utilising EEG, EOG, and EMG. The subset consists solely of patients achieving an MF1 of >35.0% when training on the whole dataset.

Patient	Age Group	PELOD Score	Neurological Condition	EEG		EEG, EOG, and EMG		EEG, EOG and EMG (subset)	
				Acc (%)	MF1 (%)	Acc (%)	MF1 (%)	Acc (%)	MF1 (%)
PICU001	0-2 Months	3	No	64.9	63.8	57.6	46.9	52.3	47.9
PICU002	0-2 Months	9	No	59.4	39.4	61.1	33.1		
PICU003	0-2 Months	7	Yes	41.4	27.1	50.2	30.7		
PICU004	0-2 Months	6	Yes*	84.6	67.4	35.4	32.9		
PICU005	0-2 Months	8	No	52.2	49.9	56.6	54.3	62.0	51.3
PICU006	0-2 Months	5	No	67.3	64.0	54.4	42.1	60.9	47.3
PICU007	0-2 Months	2	No	72.3	63.2	75.3	61.6	72.0	59.5
PICU008	0-2 Months	2	No	42.8	28.7	54.9	43.5	58.4	50.0
PICU009	0-2 Months	7	Yes	66.8	61.7	65.9	52.8	64.6	56.5
PICU010	0-2 Months	5	No	64.7	59.4	65.6	59.1	48.1	36.0
PICU011	0-2 Months	7	No	31.9	30.6	44.3	34.1		
PICU012	2-6 Months	9	No	58.4	24.6	50.6	39.3	57.8	38.8
PICU013	2-6 Months	11	No	68.1	37.9	77.4	54.4	52.1	39.2
PICU014	2-6 Months	7	Yes*	60.1	54.7	67.2	52.2	77.7	67.7
PICU015	2-6 Months	9	No	62.3	37.6	62.3	39.0	55.8	37.1
PICU016	2-6 Months	14	Yes*	81.7	61.8	65.0	50.3	67.1	50.0
PICU017	2-6 Months	9	No	68.8	55.7	73.1	56.5	77.5	62.0
PICU018	2-6 Months	4	No	74.3	30.1	83.1	57.4	85.8	60.4
PICU019	6-12 Months	21	No	60.0	32.6	58.1	29.8		
PICU020	6-12 Months	3	No	53.2	23.5	83.7	60.1	82.8	62.1
PICU021	1-3 Years	7	No	84.7	38.7	85.7	43.2	78.6	41.2
PICU022	1-3 Years	7	Yes	56.3	42.8	71.1	48.9	71.0	49.1
PICU023	5-9 Years	7	No	39.4	24.9	58.9	34.8		
PICU024	9-13 Years	7	Yes	49.4	24.2	86.8	58.9	64.8	41.8
PICU025	13-18 Years	8	Yes*	77.2	38.1	85.0	37.4	88.6	43.0
PICU026	13-18 Years	9	Yes*	33.8	24.8	70.8	55.2	74.8	52.3
PICU027	13-18 Years	5	Yes	23.5	22.9	48.7	43.7	48.4	53.9
PICU028	13-18 Years	9	Yes*	78.0	29.9	24.7	16.4		

PELOD = paediatric logistic organ dysfunction, EEG = electroencephalogram, EOG = electrooculogram, EMG = electromyogram, Acc = accuracy, MF1 = macro f1-score, * : encephalopathy.

S.14 Performance of the CNN with Inter-Patient Training per Patient of the PICU Dataset

Table 25: Comparison of performance metrics with and without inter-patient training for three-state performance of the convolutional neural network on patient level Results were obtained by 5-fold cross-validation with and without subsequent inter-patient training of 240 or 480 epochs.

Patient	No inter-patient training		Inter-patient training with 240 epochs (2 hours)		Inter-patient training with 480 epochs (4 hours)	
	Acc (%)	MF1 (%)	Acc (%)	MF1 (%)	Acc (%)	MF1 (%)
PICU001	64.9	63.8	67.8	64.5	70.3	70.1
PICU002	59.4	39.4	62.8	32.8	59.8	33.6
PICU003	41.4	27.1	71.7	47.6	72.4	48.1
PICU004	84.6	67.4	81.5	42.5	86.7	57.4
PICU005	52.2	49.9	62.5	59.4	59.8	57.0
PICU006	67.3	64.0	75.8	73.3	71.7	70.2
PICU007	72.3	63.2	77.4	56.7	76.1	70.9
PICU008	42.8	28.7	44.4	39.2	44.5	33.0
PICU009	66.8	61.7	70.9	63.0	67.6	60.7
PICU010	64.7	59.4	69.2	51.5	66.3	49.8
PICU011	31.9	30.6	62.1	37.9	59.6	34.4
PICU012	58.4	24.6	57.3	24.3	56.7	24.1
PICU013	68.1	37.9	78.8	35.9	79.0	45.1
PICU014	60.1	54.7	71.1	64.8	79.3	65.6
PICU015	62.3	37.6	85.3	56.8	86.2	57.8
PICU016	81.7	61.8	74.8	57.9	87.6	60.8
PICU017	68.8	55.7	74.7	53.9	77.7	61.9
PICU018	74.3	30.1	81.6	30.0	87.6	31.1
PICU019	60.0	32.6	83.6	63.7	90.2	80.8
PICU020	53.2	23.5	51.3	32.9	52.6	27.2
PICU021	84.7	38.7	87.7	33.4	88.7	31.8
PICU022	56.3	42.8	61.4	33.4	80.1	57.4
PICU023	39.4	24.9	36.7	18.6	40.7	24.5
PICU024	49.4	24.2	52.6	38.1	50.8	37.5
PICU025	77.2	38.1	90.3	32.2	93.6	72.7
PICU026	33.8	24.8	76.5	51.6	86.5	58.9
PICU027	23.5	22.9	85.4	57.1	88.2	58.9
PICU028	78.0	29.9	91.8	53.9	91.0	51.5

Acc = accuracy, MF1 = macro f1-score.

S.15 Overview of Manually Selected Features

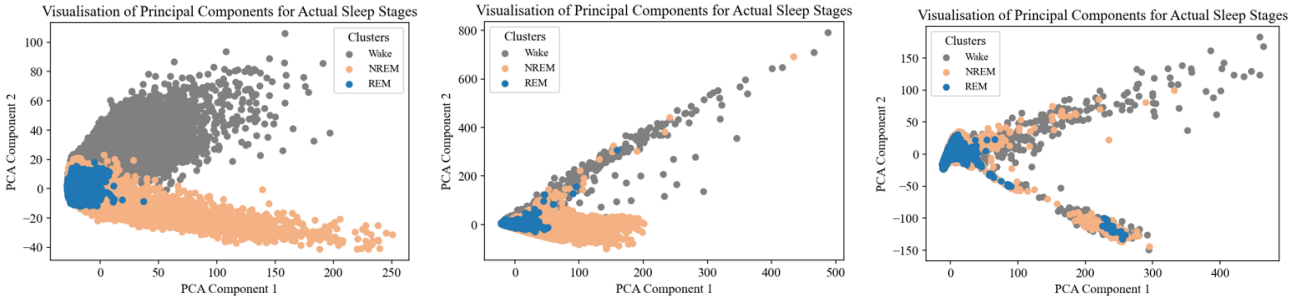
These features are utilised for the fully unsupervised approach and were described in detail by Hiemstra et al. [31]. They were calculated for each epoch.

Table 26: Manually selected electroencephalography feature categories and descriptions. Breakdown of time, frequency, and time-frequency domain features.

Feature category	Number of features	Feature description
Time domain	14	Statistical features: Mean of absolute amplitude, variance, zero-crossing-rate, interquartile range (25^{th}-75^{th}), signal sum, energy, kurtosis, skewness, Shannon entropy.
		Hjorth parameters: Activity, Mobility, Complexity.
		Higuchi fractal dimension.
		Detrended fluctuation analysis.
Frequency domain	25	Spectral bandpowers: total signal power, delta, theta, alpha, beta, gamma (relative and absolute).
		Spectral bandpower ratio: gamma/delta, gamma/theta, beta/delta, beta/theta, alpha/delta, alpha/theta.
		Sleep spindles: spectral bandpower 11-15 Hz (sigma).
		Spectral descriptors: spectral edge 95%, median and mean frequency, spectral kurtosis, spectral skewness, spectral entropy.
Time-frequency domain	12	Mean absolute value and standard deviation of coefficient amplitudes in D1, D2, D3, D4, D5 and A5 bands.

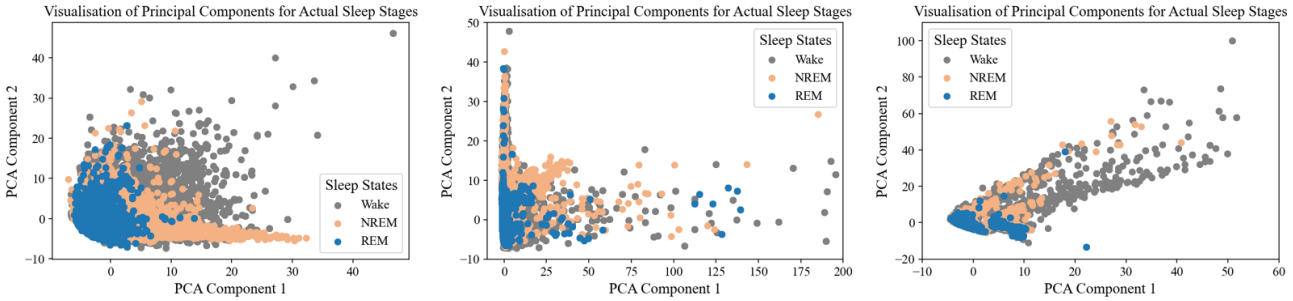
S.16 Visualisation of Principal Components

These figures illustrate the distribution of the first two principal components of either manually selected features or features from the CNN, where the different colours indicate the different sleep stages. Easily separable clusters imply distinct patterns in the feature space that correspond well to the different sleep stages. In contrast, overlapping clusters imply that the features do not clearly distinguish between sleep stages, indicating a lack of separability in the data.



(a) Sleep-EDF dataset **(b)** PSG dataset **(c)** PICU dataset

Figure 27: Visualisation of the two first principal components of the features from the convolutional neural network for the Sleep-EDF, PSG and PICU dataset. The results were obtained by a principal component analysis on each whole dataset. The different colours represent the manually labelled sleep stages.



(a) Sleep-EDF dataset **(b)** PSG dataset **(c)** PICU dataset

Figure 28: Visualisation of the two first principal components of the manually selected features for the Sleep-EDF, PSG and PICU dataset. The results were obtained by a principal component analysis on each whole dataset. The different colours represent the manually labelled sleep stages.

S.17 Distribution of Principal Components of Manually Selected Feature

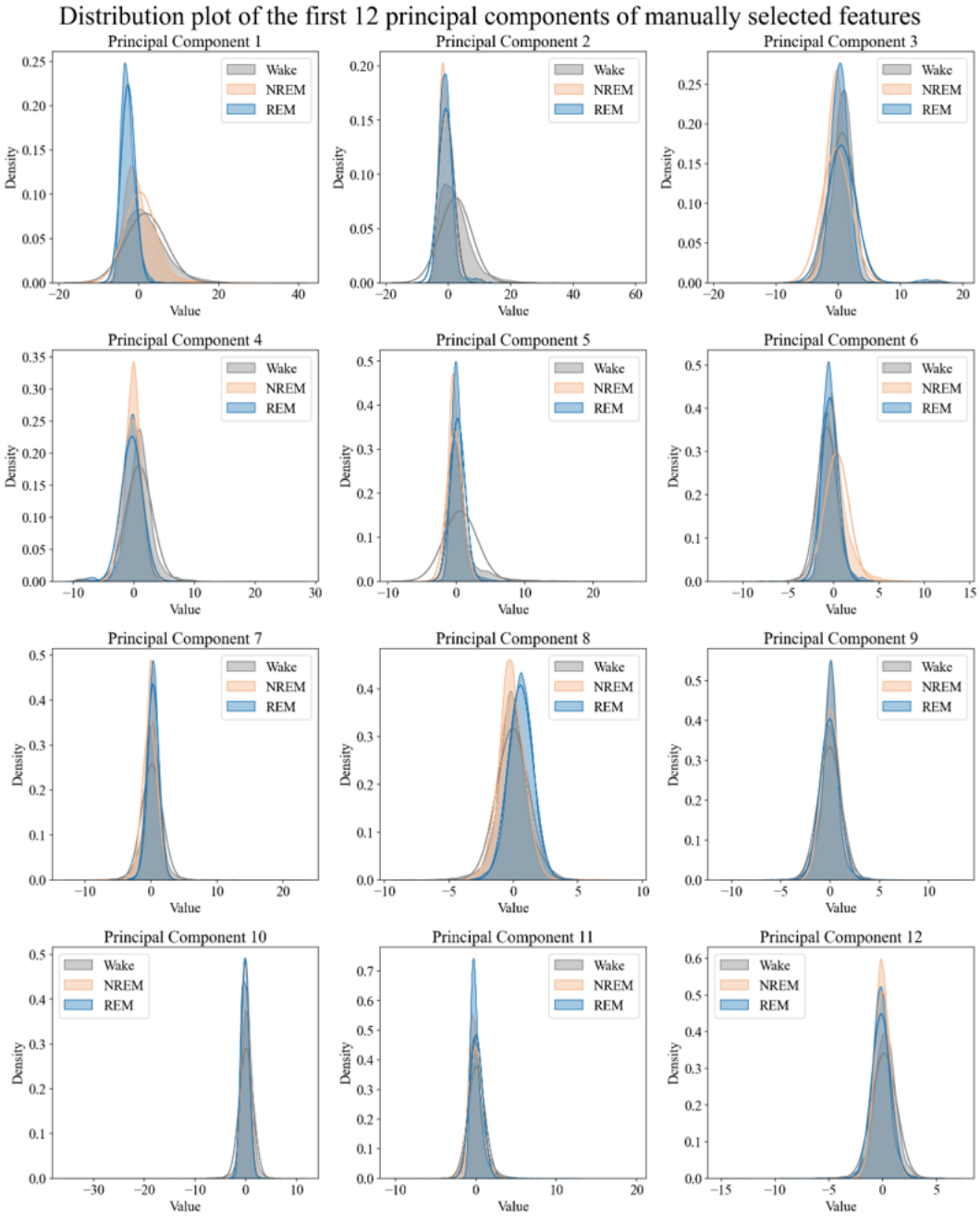


Figure 29: Distribution plots of the first 12 principal components of manually selected features from the Sleep-EDF dataset, stratified by sleep stage (Wake, NREM, REM). For each principal component, the coloured areas represent the actual distributions of the data, while the overlaid lines indicate fitted Gaussian distributions per sleep stage. Both reflect probability density functions. These plots illustrate whether features are normally distributed and the degree of overlap and separability between stages in the reduced feature space.

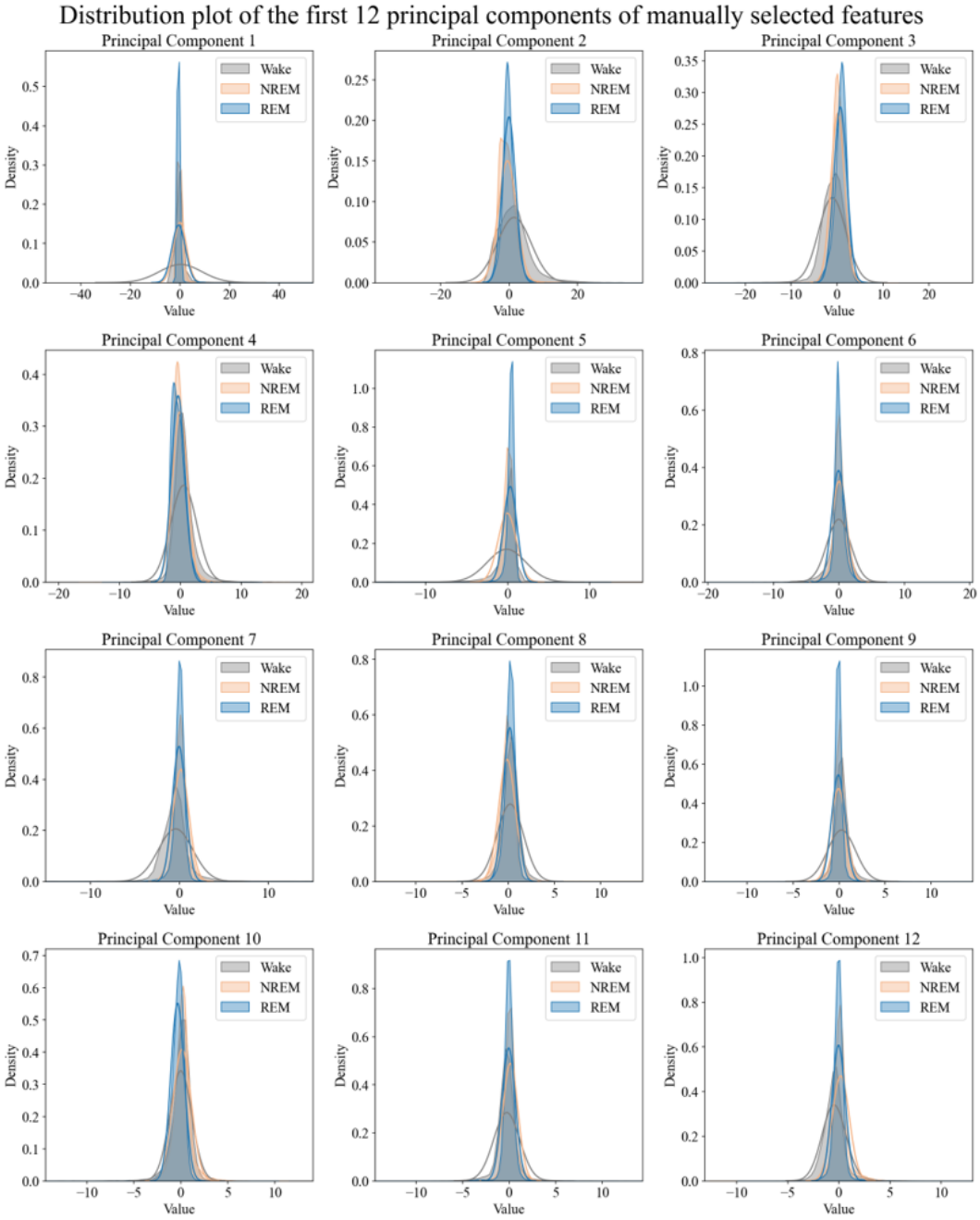


Figure 30: Distribution plots of the first 12 principal components of manually selected features from the PSG dataset, stratified by sleep stage (Wake, NREM, REM). For each principal component, the coloured areas represent the actual distributions of the data, while the overlaid lines indicate fitted Gaussian distributions per sleep stage. Both reflect probability density functions. These plots illustrate whether features are normally distributed and the degree of overlap and separability between stages in the reduced feature space.

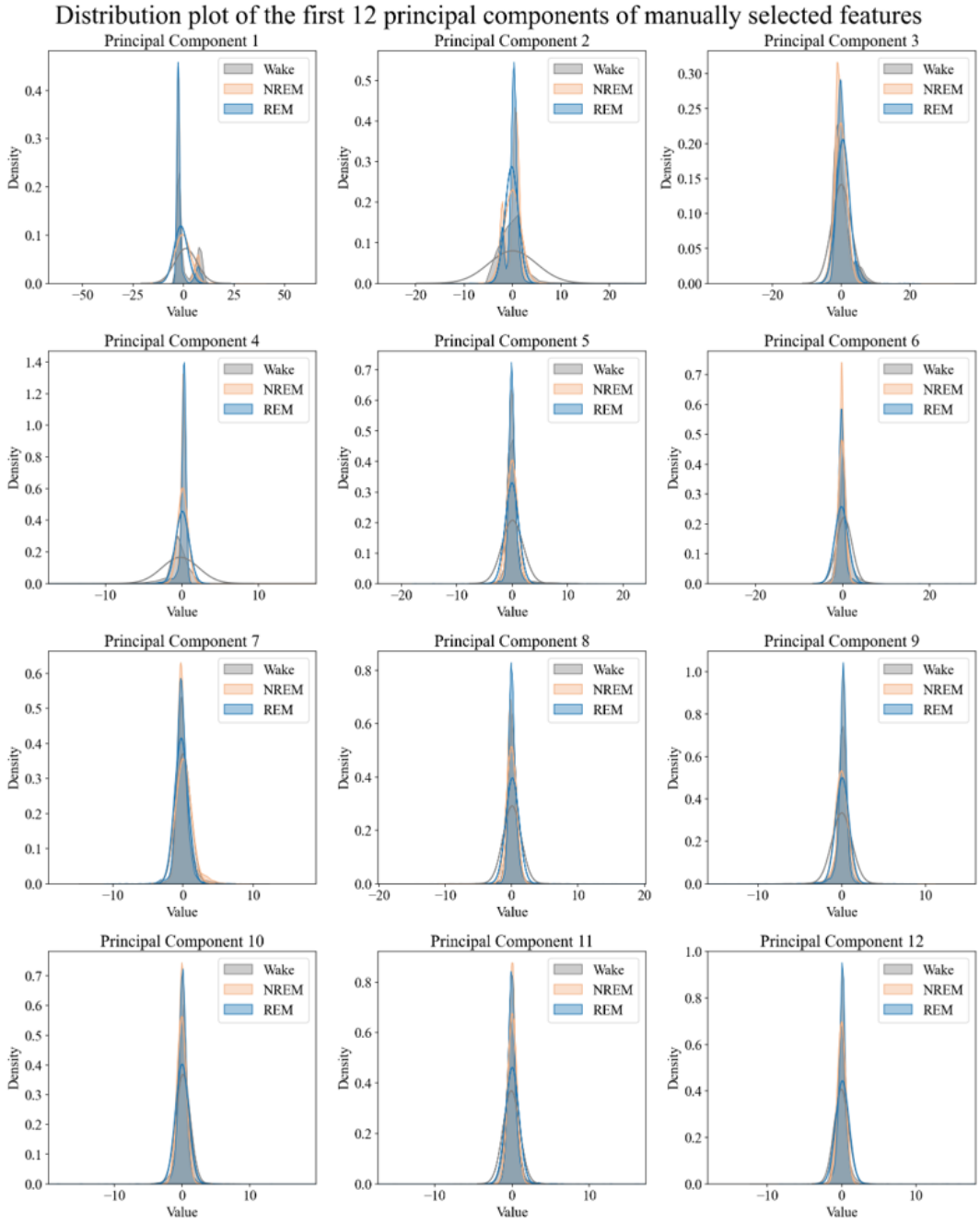
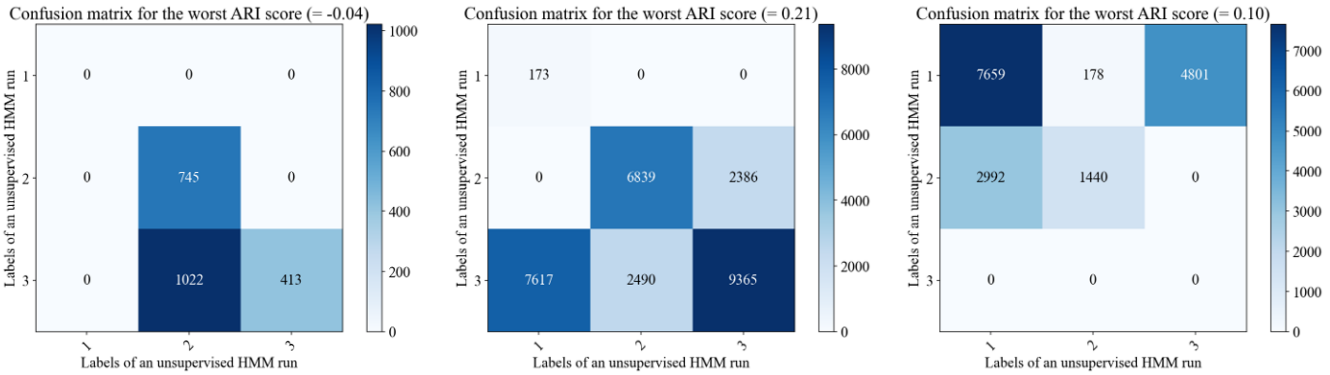


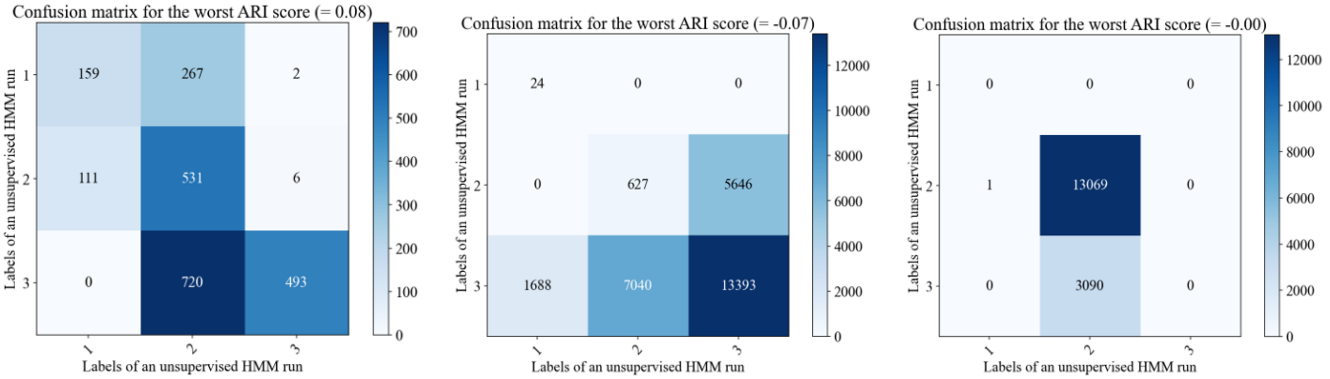
Figure 31: Distribution plots of the first 12 principal components of manually selected features from the PICU dataset, stratified by sleep stage (Wake, NREM, REM). For each principal component, the coloured areas represent the actual distributions of the data, while the overlaid lines indicate fitted Gaussian distributions per sleep stage. Both reflect probability density functions. These plots illustrate whether features are normally distributed and the degree of overlap and separability between stages in the reduced feature space.

S.18 Confusion Matrices of the Unsupervised HMM with the Worst ARI Scores



(a) Sleep-EDF Dataset **(b)** PSG Dataset **(c)** PICU Dataset

Figure 32: Confusion Matrices showing alignment between clusterings of two hidden Markov models utilising features from the convolutional neural network. The results are shown for the models that achieved the lowest adjusted rand index score. Cluster numbers are relabelled to match each other, potentially visualising the clusters on the diagonal.



(a) Sleep-EDF Dataset **(b)** PSG Dataset **(c)** PICU Dataset

Figure 33: Confusion Matrices showing alignment between clusterings of two hidden Markov models utilising manual features. The results are shown for the models that achieved the lowest adjusted rand index score. Cluster numbers are relabelled to match each other, potentially visualising the clusters on the diagonal.

S.19 Stability of the Unsupervised Clustering over Varying Numbers of Clusters

The following figures and tables illustrate the stability of the unsupervised clustering for the partly unsupervised approach, for a varying amount of clusters to identify. To better interpret the stability, the tables are provided, which give insight into the number of clusters that are actually identified.

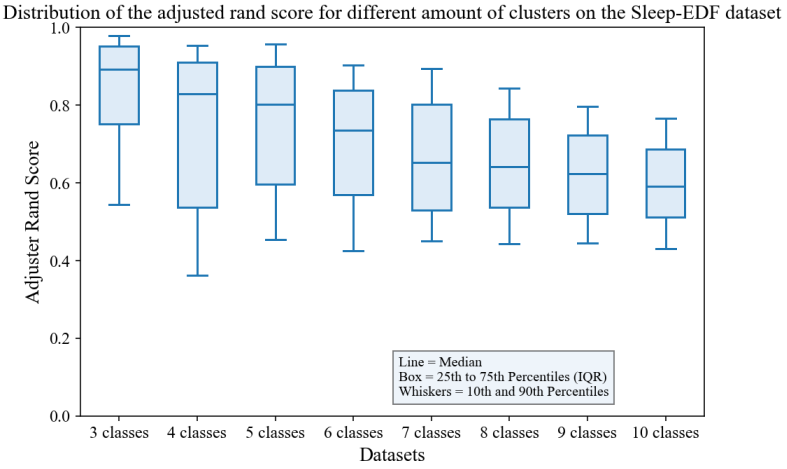


Figure 34: Boxplots comparing the clustering stability through adjusted rand index scores for different numbers of clustering in the Sleep-EDF dataset. Scores were computed between the 10 best-performing hidden Markov models, utilising features from the convolutional neural network, within each fold. The mean and percentiles are calculated over all 20 folds.

Table 27: The number of clusters classified by the unsupervised hidden Markov model for the Sleep-EDF dataset, with a varying number of clusters to identify. The results were obtained for 10 runs for each fold within the 20-fold cross validation, resulting in 200 runs.

	Number (%) of model runs classifying each cluster count									
Input number of clusters	**One**	**Two**	**Three**	**Four**	**Five**	**Six**	**Seven**	**Eight**	**Nine**	**Ten**
3 clusters	0 (0)	10 (5)	190 (95)	-	-	-	-	-	-	-
4 clusters	0 (0)	7 (4)	18 (9)	175 (88)	-	-	-	-	-	-
5 clusters	0 (0)	0 (0)	10 (5)	27 (14)	163 (82)	-	-	-	-	-
6 clusters	0 (0)	3 (2)	7 (4)	4 (2)	30 (15)	156 (78)	-	-	-	-
7 clusters	0 (0)	0 (0)	9 (5)	7 (4)	12 (6)	31 (16)	141 (71)	-	-	-
8 clusters	0 (0)	0 (0)	1 (1)	10 (5)	12 (6)	18 (9)	70 (35)	89 (45)	-	-
9 clusters	0 (0)	0 (0)	8 (4)	2 (1)	11 (6)	13 (7)	30 (15)	72 (36)	64 (32)	-
10 clusters	0 (0)	0 (0)	0 (0)	10 (5)	9 (5)	11 (6)	7 (4)	46 (23)	86 (43)	31 (16)

.

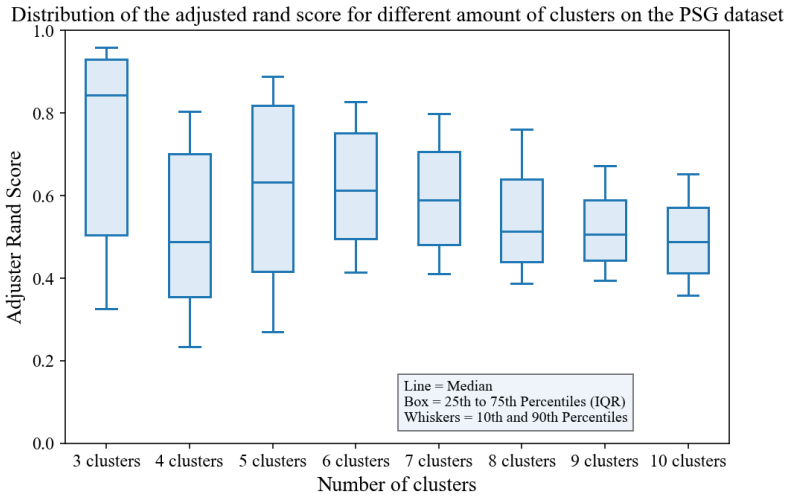


Figure 35: Boxplots comparing the clustering stability through adjusted rand index scores for different numbers of clustering in the PSG dataset. Scores were computed between the 10 best-performing hidden Markov models, utilising features from the convolutional neural network, within each fold. The mean and percentiles are calculated over all 20 folds.

Table 28: The number of clusters classified by the unsupervised hidden Markov model for the PSG dataset, with a varying number of clusters to identify. The results were obtained for 10 runs for each fold within the 5-fold cross-validation, resulting in 50 runs per row.

	Number (%) of model runs classifying each cluster count									
Input number of clusters	**One**	**Two**	**Three**	**Four**	**Five**	**Six**	**Seven**	**Eight**	**Nine**	**Ten**
3 clusters	0 (0)	0 (0)	50 (100)	-	-	-	-	-	-	-
4 clusters	0 (0)	0 (0)	0 (0)	50 (100)	-	-	-	-	-	-
5 clusters	0 (0)	0 (0)	0 (0)	2 (4)	48 (96)	-	-	-	-	-
6 clusters	0 (0)	0 (0)	0 (0)	0 (0)	10 (20)	40 (80)	-	-	-	-
7 clusters	0 (0)	0 (0)	0 (0)	0 (0)	2 (4)	12 (24)	36 (72)	-	-	-
8 clusters	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	3 (6)	16 (32)	31 (62)	-	-
9 clusters	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	8 (16)	11 (22)	31 (62)	-
10 clusters	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (2)	4 (8)	35 (70)	10 (20)

.

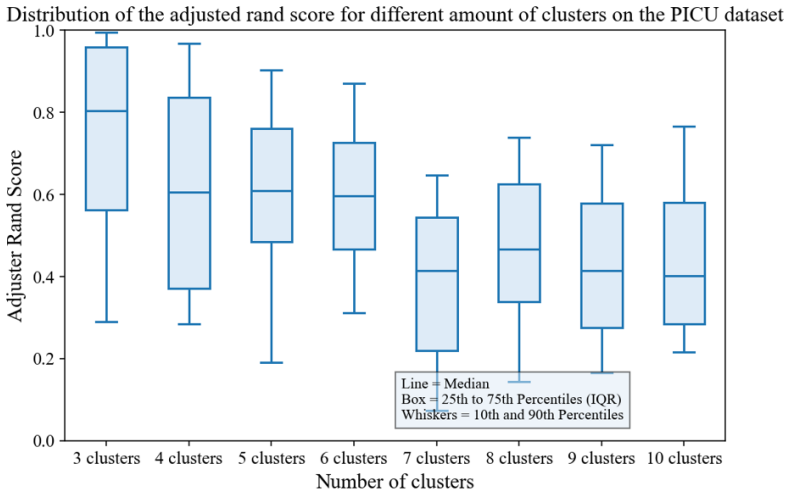


Figure 36: Boxplots comparing the clustering stability through adjusted rand index scores for different numbers of clustering in the PICU dataset. Scores were computed between the 10 best-performing hidden Markov models, utilising features from the convolutional neural network, within each fold. The mean and percentiles are calculated over all 20 folds.

Table 29: The number of clusters classified by the unsupervised hidden Markov model for the PICU dataset, with a varying number of clusters to identify. The results were obtained for 10 runs for each fold within the 5-fold cross-validation, resulting in 50 runs per row.

	Number (%) of model runs classifying each cluster count									
Input number of clusters	**One**	**Two**	**Three**	**Four**	**Five**	**Six**	**Seven**	**Eight**	**Nine**	**Ten**
3 clusters	0 (0)	18 (36)	32 (64)	-	-	-	-	-	-	-
4 clusters	0 (0)	12 (24)	24 (48)	14 (28)	-	-	-	-	-	-
5 clusters	0 (0)	11 (22)	3 (6)	27 (54)	9 (18)	-	-	-	-	-
6 clusters	0 (0)	4 (8)	7 (14)	19 (38)	13 (26)	7 (14)	-	-	-	-
7 clusters	0 (0)	1 (2)	1 (2)	11 (22)	24 (48)	9 (18)	4 (8)	-	-	-
8 clusters	0 (0)	0 (0)	0 (0)	9 (18)	18 (36)	19 (38)	4 (8)	0 (0)	-	-
9 clusters	0 (0)	0 (0)	0 (0)	2 (4)	14 (28)	19 (38)	15 (30)	0 (0)	0 (0)	-
10 clusters	0 (0)	0 (0)	0 (0)	2 (4)	18 (36)	13 (26)	8 (16)	6 (12)	3 (6)	0 (0)

S.20 Performance of the Supervised HMM Utilising Manually Selected Features

Table 30: Performance metrics for three- and five-state sleep stage classification for the supervised hidden Markov model utilising manual features Results obtained by 20- and 5-fold cross-validation for the Sleep-EDF and PSG/PICU dataset, respectively. The italicised results correspond to 92 out of 120 patients from the PSG dataset.

	Three-state performance in % (SD)				Five-state performance in % (SD)			
Dataset	**Acc**	**MF1**	**Kappa**	**AUC**	**Acc**	**MF1**	**Kappa**	**AUC**
Sleep-EDF	79.9 (6.1)	76.4 (7.2)	64.9 (10.5)	92.4 (4.5)	72.0 (8.7)	64.1 (10.0)	61.8 (11.8)	90.8 (4.6)
PSG	56.4 (7.1)	54.0 (5.3)	33.3 (6.6)	80.8 (1.4)	*44.5 (4.3)*	*36.6 (4.4)*	*26.7 (6.7)*	*76.1 (3.0)*
PICU	52.5 (10.9)	37.3 (6.5)	15.9 (14.4)	59.4 (8.8)	n/a	n/a	n/a	n/a

SD = standard deviation, Acc = accuracy, MF1 = macro-f1 score, kappa = Cohen's kappa, AUC = area under the curve

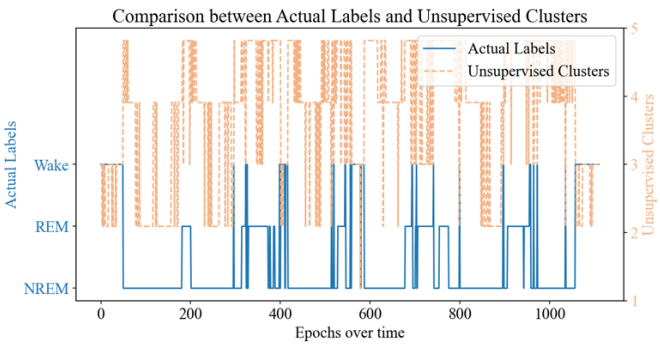
S.21 Visualisation of the Unsupervised Clusters for All Datasets

In this appendix three types of visualisations are shown:

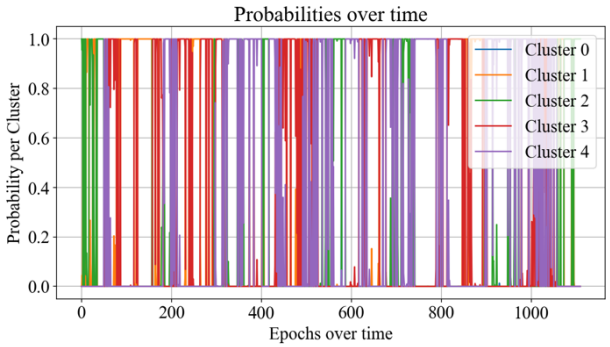
- **Hypnograms over time:** A Hypnogram for manually assigned labels and a hypnogram for unsupervised clusters are shown over time, enabling comparison.
- **Probabilities over time:** The probability of each unsupervised cluster over time is represented on a scale from 0 to 1, where a higher probability indicates a greater likelihood that the corresponding time point belongs to that cluster.
- **Visualisation of Principal Components:** The first two principal components are visualised, where the colours represent either the manually labelled sleep stages or the unsupervised cluster. This again enables comparison.

S.21.1 Sleep-EDF Dataset

Partly Unsupervised Approach

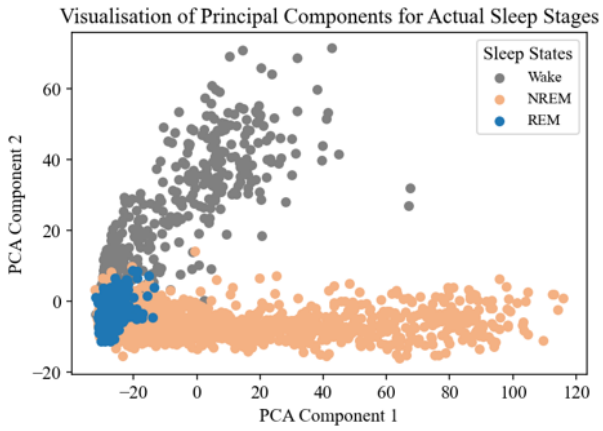


(a) Hypnogram over time.

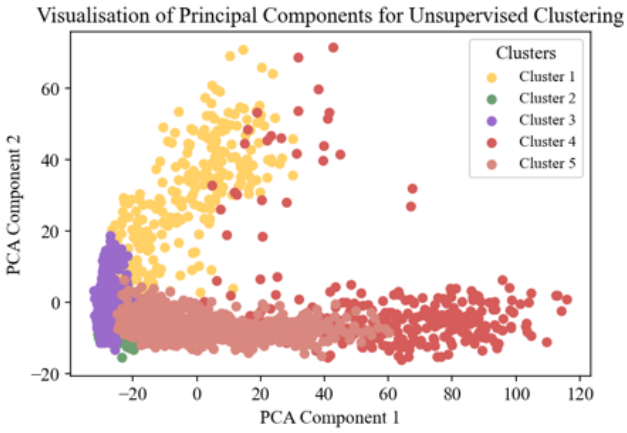


(b) Probabilities over time.

Figure 37: Visualisation of partly unsupervised clusters for a whole night of sleep a patient from the Sleep-EDF dataset. Clusters (a) and their corresponding probabilities (b) are obtained from an unsupervised hidden Markov model utilising features from the convolutional neural network.



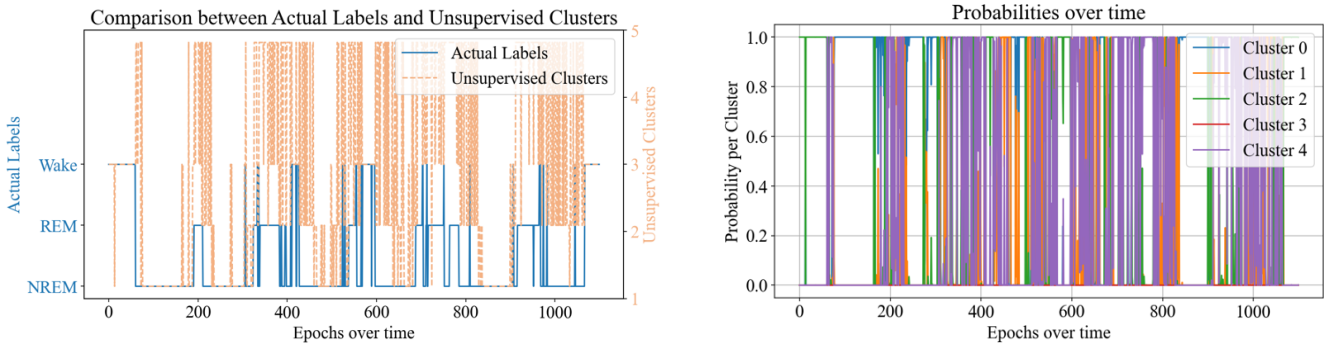
(a) Manually labelled sleep stages



(b) Unsupervised clusters

Figure 38: Comparison between manually labelled sleep stages and partly unsupervised clusters through visualisation of the first two principal components for a patient from the Sleep-EDF dataset. Clusters were obtained from the unsupervised hidden Markov model utilising features from the convolutional neural network.

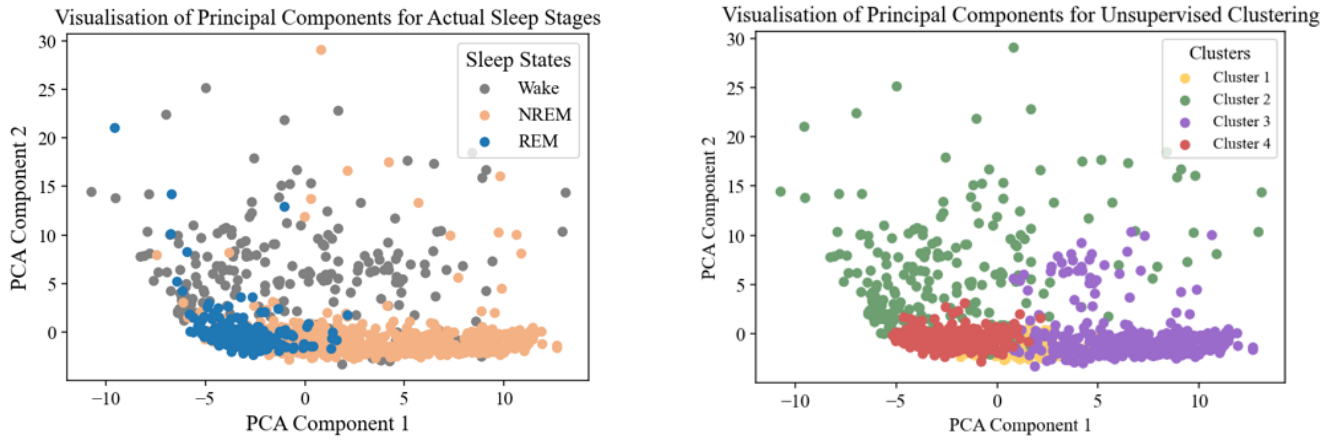
Fully Unsupervised Approach



(a) Hypnogram over time.

(b) Probabilities over time.

Figure 39: Visualisation of fully unsupervised clusters for a whole night of sleep a patient from the Sleep-EDF dataset. Clusters (a) and their corresponding probabilities (b) are obtained from an unsupervised hidden Markov model utilising manually selected features.



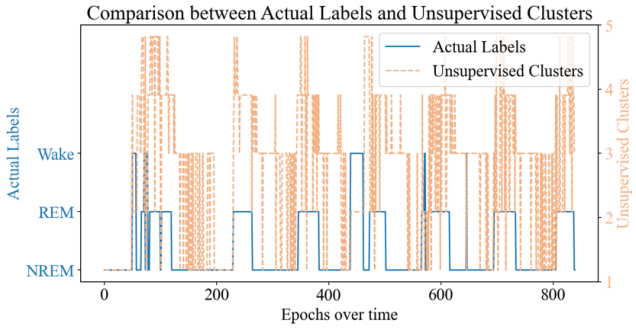
(a) Manually labelled sleep stages

(b) Unsupervised clusters

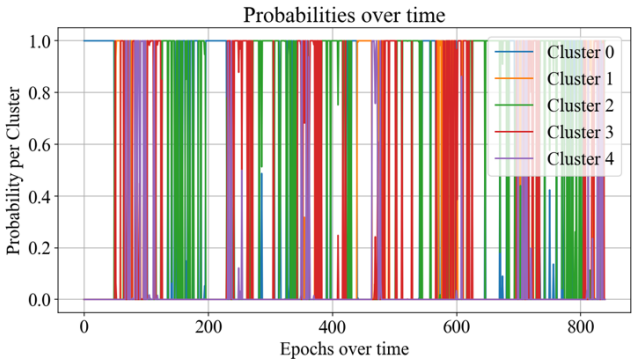
Figure 40: Comparison between manually labelled sleep stages and fully unsupervised clusters through visualisation of the first two principal components for a patient from the Sleep-EDF dataset. Clusters were obtained from the unsupervised hidden Markov model utilising manually selected features.

S.21.2 PSG Dataset

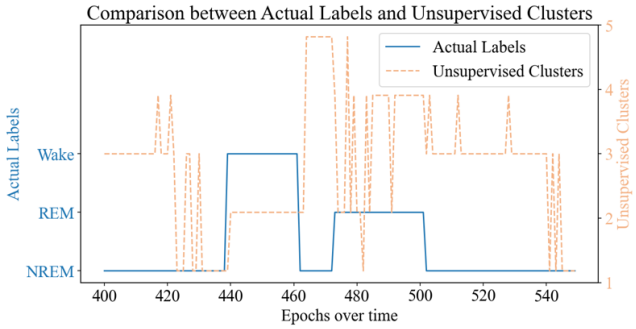
Partly Unsupervised Approach



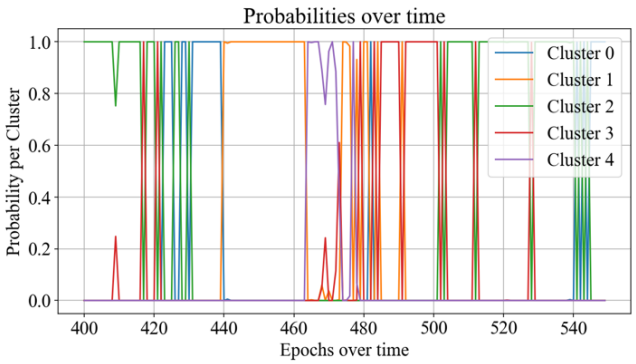
(a) Hypnogram - full night



(b) Probabilities - full night

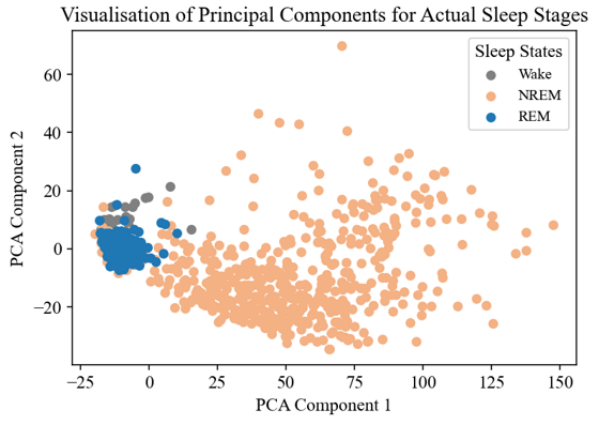


(c) Hypnogram - selected time period

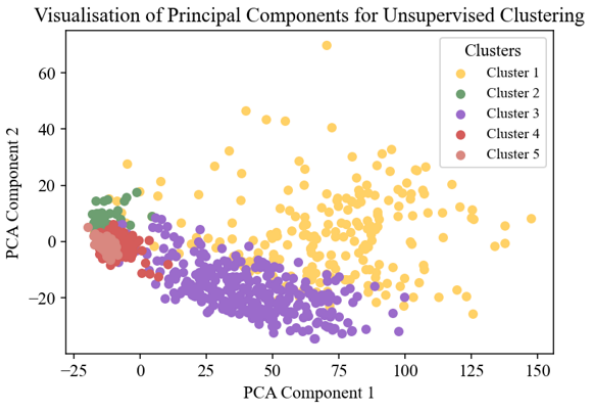


(d) Probabilities - selected time period

Figure 41: Visualisation of partly unsupervised clusters for a patient from the PSG dataset. Subfigures (a) and (b) show the full night; (c) and (d) display a selected time window. Clusters and their probabilities were obtained from an unsupervised hidden Markov model utilising features from the convolutional neural network.



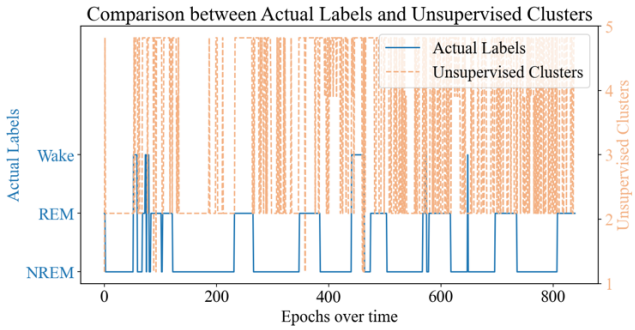
(a) Manually labelled sleep stages



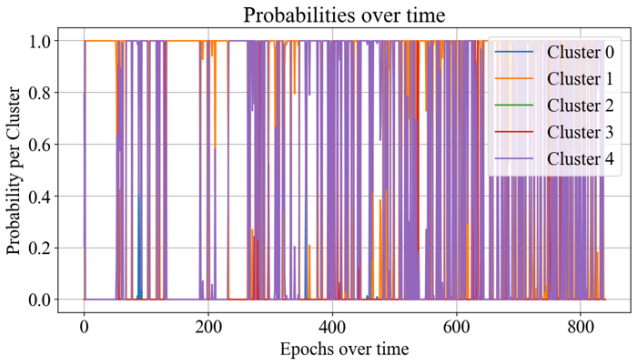
(b) Unsupervised clusters

Figure 42: Comparison between manually labelled sleep stages and partly unsupervised clusters through visualisation of the first two principal components for a patient from the PSG dataset. Clusters were obtained from the unsupervised hidden Markov model utilising features from the convolutional neural network.

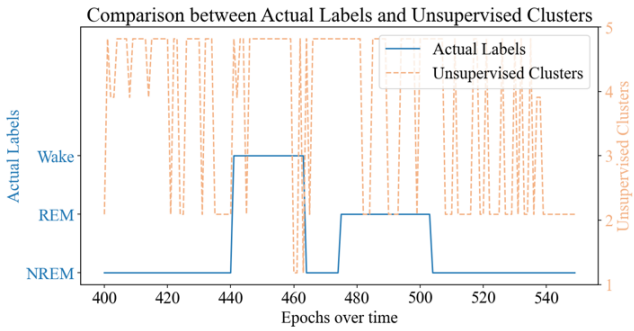
Fully Unsupervised Approach



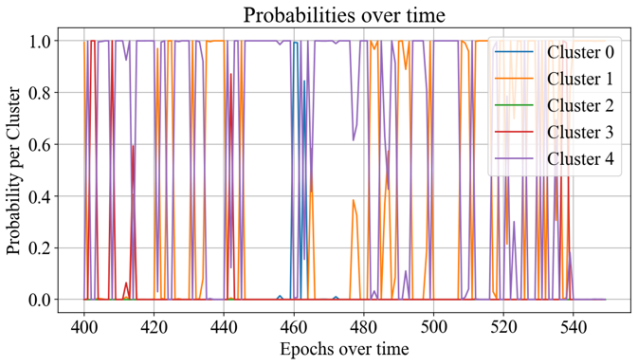
(a) Hypnogram - full night



(b) Probabilities - full night

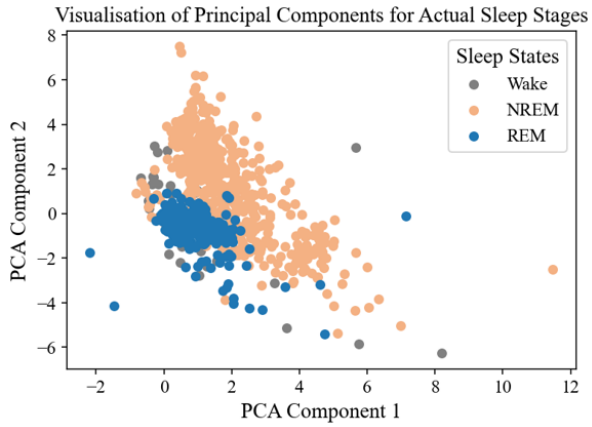


(c) Hypnogram - selected time period

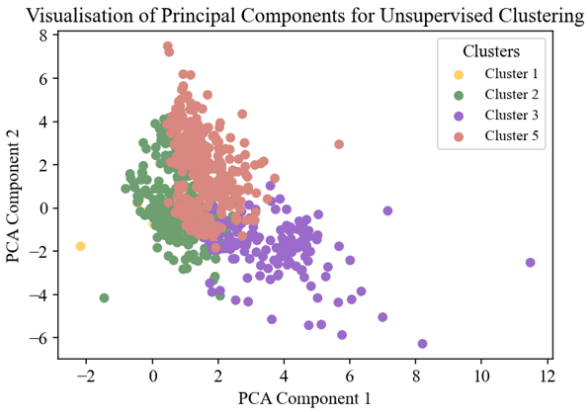


(d) Probabilities - selected time period

Figure 43: Visualisation of fully unsupervised clusters for a patient from the PSG dataset. Subfigures (a) and (b) show the full night; (c) and (d) display a selected time window. Clusters and their probabilities were obtained from an unsupervised hidden Markov model utilising manually selected features.



(a) Manually labelled sleep stages

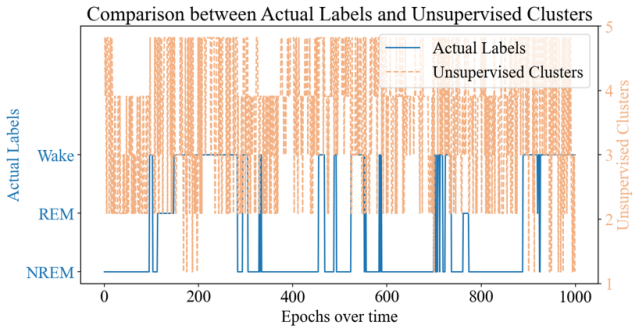


(b) Unsupervised clusters

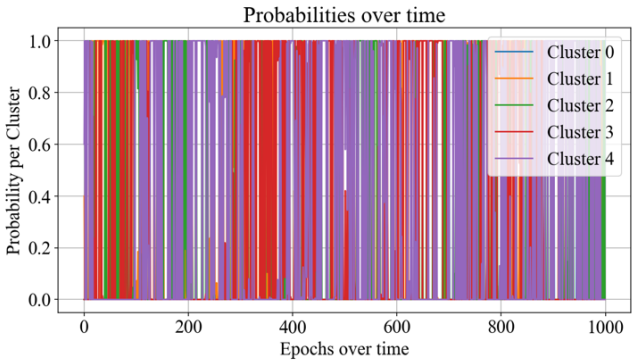
Figure 44: Comparison between manually labelled sleep stages and fully unsupervised clusters through visualisation of the first two principal components for a patient from the PSG dataset. Clusters were obtained from the unsupervised hidden Markov model utilising manually selected features.

S.21.3 PICU dataset

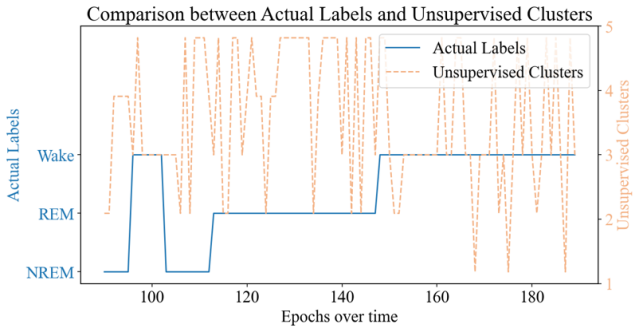
Partly Unsupervised Approach



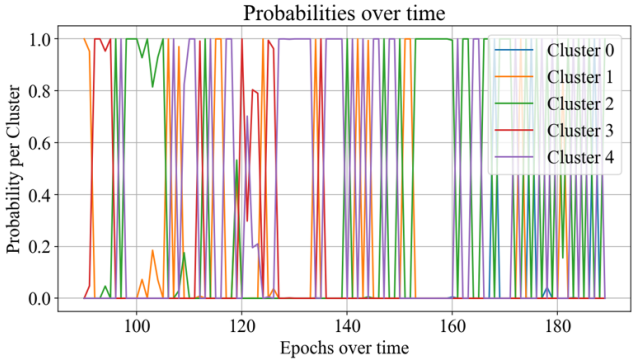
(a) Hypnogram - full night



(b) Probabilities - full night

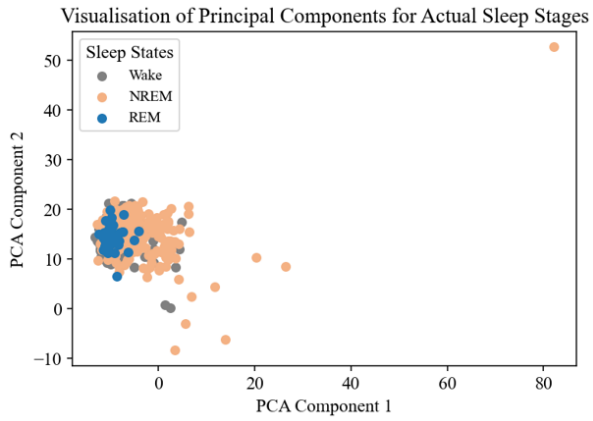


(c) Hypnogram - selected time period

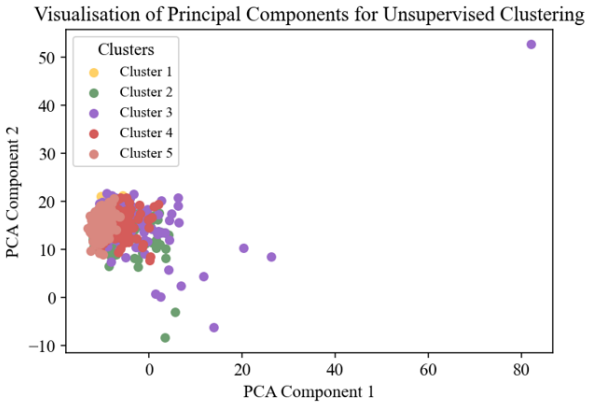


(d) Probabilities - selected time period

Figure 45: Visualisation of partly unsupervised clusters for a patient from the PICU dataset. Subfigures (a) and (b) show the full night; (c) and (d) display a selected time window. Clusters and their probabilities were obtained from an unsupervised hidden Markov model utilising features from the convolutional neural network.



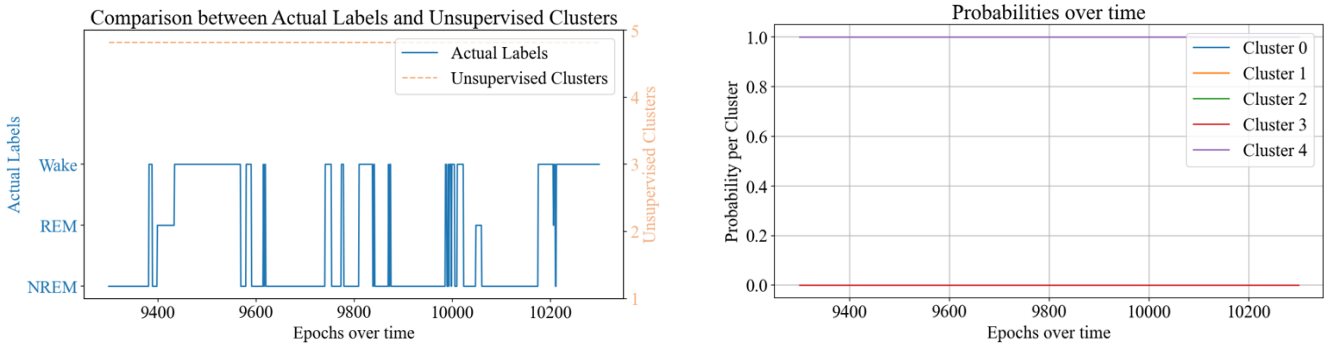
(a) Manually labelled sleep stages



(b) Unsupervised clusters

Figure 46: Comparison between manually labelled sleep stages and partly unsupervised clusters through visualisation of the first two principal components for a patient from the PICU dataset. Clusters were obtained from the unsupervised hidden Markov model utilising features from the convolutional neural network.

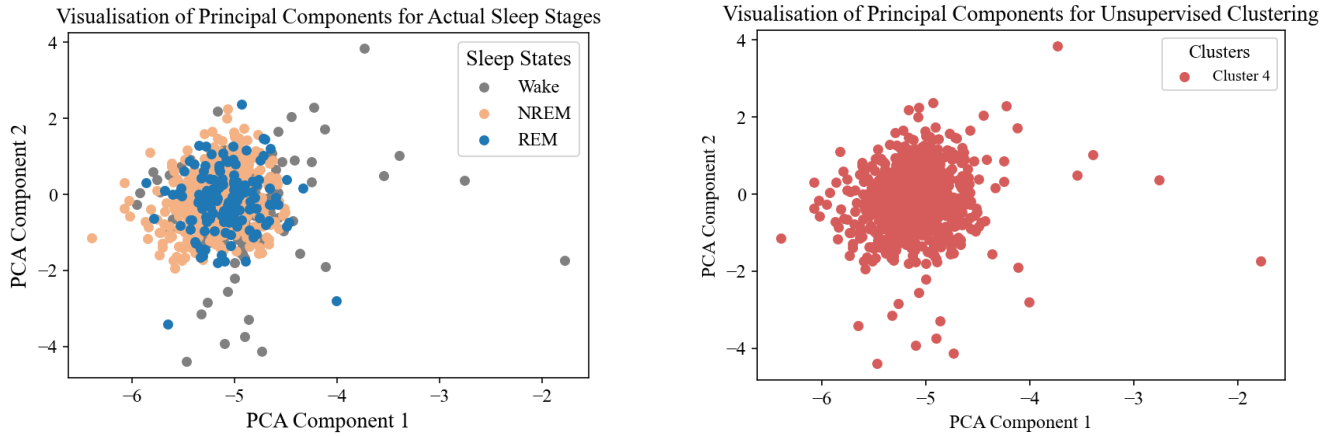
Fully Unsupervised Approach



(a) Hypnogram over time.

(b) Probabilities over time.

Figure 47: Visualisation of fully unsupervised clusters for a whole night of sleep a patient from the PICU dataset. Clusters (a) and their corresponding probabilities (b) are obtained from an unsupervised hidden Markov model utilising manually selected features.



(a) Manually labelled sleep stages

(b) Unsupervised clusters

Figure 48: Comparison between manually labelled sleep stages and fully unsupervised clusters through visualisation of the first two principal components for a patient from the PICU dataset. Clusters were obtained from the unsupervised hidden Markov model utilising manually selected features.