Infant Sleep Staging with Wearable Sensors and Deep Learning: Evaluating and Controlling Age-Related Effects

Matias Häggman

School of Science

Bachelor's thesis
Espoo 13/10/2024

Supervisor

Prof. Fabricio Oliveira

Advisor

DSc. (Tech.) Manu Airaksinen

Aalto University
School of Science

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the bachelor's thesis

Author Matias Häggman

Title Infant Sleep Staging with Wearable Sensors and Deep Learning: Evaluating and Controlling Age-Related Effects

Degree programme Engineering Physics and Mathematics

Major Mathematics and Systems Sciences **Code of major** SCI3029

Teacher in charge Prof. Fabricio Oliveira

Advisor DSc. (Tech.) Manu Airaksinen

Date 13/10/2024 **Number of pages** 47 **Language** English

Abstract

Sleep problems in early childhood have been linked to a wide range of health and developmental problems. Sleep staging is used in clinical studies to diagnose problems with sleep health. Standardized sleep staging requires a medical study of sleep, usually polysomnography, including an electroencephalography (EEG) analysis by a medical professional. For this reason, standard sleep staging is not a viable solution for long-term studies of sleep at-home, and alternative methods are needed that allow objective monitoring of sleep in the child's natural sleep environment.

Ranta et al. (2021) have developed NApping PAnts (NAPPA), a wearable sensor for infants that uses an accelerometer and a gyroscope to record abdominal movements during sleep. We have developed a Bidirectional Gated Recurrent Unit -based deep learning classifier for automatic sleep staging of infant sleep with the NAPPA wearable using an annotated dataset of 36 recordings. The classifier achieved a median accuracy of 77% in detecting stages of deep sleep, light sleep, and wake using five movement activity and respiration related features derived from the sensor signal. Infant sleep and its indicators, such as respiration rate, are dynamic and change throughout the first two years after birth. This variability in sleep stage indicators introduces potential challenges for deep learning-based classifiers, as the data used to train them are usually composed of infants of varying ages.

In this thesis, we carry out a data analysis focusing on studying how the varying age of infants impacts the features recorded by the NAPPA system, and subsequently we attempt to find methods to refine the classifier's performance in sleep staging, addressing the dynamic nature of infant sleep patterns and the impact of age on sleep behavior.

We found a distinct relationship between the age of infants and recorded features and age and model performance. Our proposed methods, however, did not ultimately offer any improvement in model performance. Despite the lack of improvement in performance, our findings highlight the role of age in the context of sleep staging using wearable sensors in infant populations. The relationship between age, sleep patterns, and recorded features suggests that age-specific adjustments may be necessary to enhance the accuracy of sleep stage classification models.

Keywords Automatic Sleep Stage Classification, Machine Learning, Deep Learning, Gated Recurrent Unit, Wearable Sensor, Pediatric Sleep Medicine

Aalto University
School of Science

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Tekniikan kandidaatintyön tiivistelmä

Tekijä Matias Häggman

Työn nimi Vauvojen unen vaiheistus puettavien antureiden ja syväoppimisen avulla: Ikäsidonnaisten vaikutusten arviointi ja hallinta

Koulutusohjelma Teknillinen fysiikka ja matematiikka

Pääaine Matematiikka ja systeemitieteet **Pääaineen koodi** SCI3029

Vastuuopettaja ja ohjaaja Prof. Fabricio Oliveira

Päivämäärä 13/10/2024 **Sivumäärä** 47 **Kieli** Englanti

Tiivistelmä

Varhaislapsuuden uniongelmat on yhdistetty monenlaisiin terveys- ja kehitysongelmiin. Unen vaiheistusta käytetään kliinisissä tutkimuksissa unen terveyteen liittyvien ongelmien diagnosoimiseksi. Standardoitu unen vaiheistus edellyttää lääketieteellistä unitutkimusta, tavallisesti polysomnografiaa, johon kuuluu myös lääketieteen ammattilaisen tekemä elektroenkefalografian (EEG) analyysi. Tavanomainen unen vaiheistus ei ole toimiva ratkaisu unen tutkimuksiin kotioloissa, ja tarvitaan vaihtoehtoisia menetelmiä, jotka mahdollistavat unen pitkäaikaisen seurannan lapsen kotona.

Ranta et al. (2021) ovat kehittäneet vauvoille puettavan NApping PAnts (NAPPA) -laitteen, joka käyttää kiihtyvyysanturia ja gyroskooppia tallentaakseen vatsan liikkeitä unen aikana. Olemme kehittäneet kaksisuuntaiseen portitettuun toistuvaan yksikköön (eng. bidirectional gated recurrent unit) -pohjaisen syväoppimisluokittimen vauvojen unen automaattiseen unen vaiheistamiseen NAPPA laitteen avulla käyttäen annotoitua 36 tallenteen tietoaineistoa. Luokittimen mediaanitarkkuus syvän unen, kevyen unen ja heräämisen vaiheiden havaitsemisessa oli 77 % käyttäen viittä anturisignaalista johdettua liikkeen aktiivisuutta ja hengitykseen liittyviä ominaisuuksia hyödyntäen.

Vauvojen uni ja sen indikaattorit, kuten hengitystaajuus, ovat dynaamisia ja muuttuvat ensimmäisten kahden vuoden aikana syntymän jälkeen. Tämä vaihtelevuus aiheuttaa mahdollisesti haasteita syväoppimiseen perustuville luokittelijoille, sillä niiden kouluttamiseen käytettävä data koostuu yleensä eri-ikäisistä vauvoista.

Tässä tutkielmassa suoritamme data-analyysin, jossa keskitymme tutkimaan, miten vaihteleva ikä vaikuttaa NAPPA-järjestelmän tallentamiin piirteisiin, ja sen jälkeen yritämme löytää menetelmiä, joilla voidaan parantaa luokittelijan suorituskykyä, ottaen huomioon vauvojen unen dynaamisen luonteen ja iän vaikutuksen.

Löysimme selvän yhteyden vauvojen iän ja tallennettujen piirteiden sekä iän ja luokittimen suorituskyvyn välillä. Ehdotetut menetelmämme eivät kuitenkaan lopulta parantaneet luokittelijan suorituskykyä. Vaikka suorituskyky ei parantunut, havaintomme korostavat iän merkitystä vauvojen automaattisen unen vaiheistuksen yhteydessä, kun käytetään puettavia antureita. Iän, unitottumusten ja tallennettujen piirteiden välinen suhde viittaa siihen, että ikäkohtaiset mukautukset saattavat olla hyödyksi uniluokittelijoiden tarkkuuden parantamisessa.

Avainsanat Automatisoitu univaiheiden luokittelu, Koneoppiminen, Syväoppiminen, Portitettu toistuva yksikkö, Puettava anturi, Lasten unilääketiede

Contents

1 Introduction

Sleep problems in infancy and early childhood have been linked to a wide range of health and developmental problems, including weight gain, increased frequency of common diseases, and learning difficulties (Liu et al., 2024). The immediate health effects of sleep deprivation include increased fatigue, depressed mood, and a weak immune response (Mervaala et al., 2019, p. 223). Research suggests that 20 - 30% of children, regardless of age, experience sleep disorders (Bruni and Novelli, 2010, Tham et al., 2017). Therefore, identifying and treating the causes of sleep disorders in early childhood and infancy is essential for general well-being and the promotion of healthy growth and development in children.

Identifying sleep disorders and disturbances of sleep in children requires a medical study of sleep. Standard methods to study sleep in children include sleep diary and questionnaires for long-term studies and polysomnography (PSG) study for short-term studies. Actigraphy devices worn by the ankle or wrist that monitor body movements are also used. In a polysomnography study, multiple physiological signals are measured and recorded in a sleep laboratory in a hospital. A sleep medicine professional manually analyzes the signals and identifies the stages of sleep (known as scoring or sleep staging) and possible underlying problems with sleep health. Although PSG is a reliable method for assessing sleep quality and diagnosing sleep disorders with an inter-rater agreement between medical professionals ranging from 0.71 to 0.81 (Cohen's kappa; 95% CI) (Lee et al., 2022), it is expensive and the hospital sleep laboratory environment is not a natural sleep environment for infants and children. On the other hand, questionnaires and sleep diaries are not based on objective physiological measurements, but rather on qualitative judgement of parents. Since parents' judgments can be subjective with substantial inter-rater differences, sleep diaries and questionnaires do not provide an objective approach in long-term studies. Lastly, traditional limb actigraphy is limited by the fact that it does not provide data on respiration or sleep stages, and therefore cannot provide crucial information on sleep quality and the sleep-wake cycle. (Sadeh, 2015)

Given the limitations of standard methods of sleep assessment in infants and children, there is a need for alternative methods that allow objective monitoring of sleep behavior in long-term out-of-hospital settings where monitoring can be performed in the child's natural sleep environment. In light of this, Ranta et al. introduced NAPPA (NApping PAnts, Ranta et al. (2021a)), a wearable device designed for infants that employs a programmable, detachable movement sensor which is worn on the front of the diaper. The sensor combines an accelerometer and a gyroscope to measure abdominal movements, and from the movement sensor signal various sleep-relevant respiratory features such as movement activity, respiration rate and respiration stability are computed. This system allows for the tracking of infant sleep with minimal disturbance, is low cost, and has been proven to reliably and accurately measure respiratory movements. Combined with the clinical data collected from PSG studies, the sensor data can be used to train a machine learning based classifier to automatically identify different sleep stages. The trained classifier can be in turn used in automatic sleep staging in nonclinical environments for long-term assessment of sleep quality and problems in infants.

For the purpose of automatic classification of sleep stages using the NAPPA system, we have developed a Bidirectional Gated Recurrent Unit (BiGRU) (Cho et al., 2014), a type of recurrent neural network model. This model initially achieved a good overall classification accuracy of 77% on a dataset consisting of 36 individual sleep recordings from distinct infants. The performance of the model was further validated in a study conducted by de Sena et al. (2024).

However, it is known that infant sleep and the sleep related physiological characteristics are highly dynamic and change considerably throughout the first two years of development after birth (Patel et al., 2022). For this reason, we identified the need to study whether the classifier performance could be further improved in this population by incorporating age-specific adjustments.

In this thesis, we seek to enhance the performance of the BiGRU classifier by using two separate modeling techniques. First, by using each infant's age as a separate input feature in the BiGRU model, we aim to teach the neural network to independently correct for the age based differences in the data. In the second approach we instead apply age-based transformations to the sensor data as a way of manually correcting for the age related differences.

The structure and content of this thesis are as follows: The background section (2) delves into the physiology and development of infant sleep, describes current established methods that are used in medical studies of sleep, and introduces the findings of previous research on automatic sleep stage classification. The details of the applied machine learning methods, such as the working principles of recurrent neural networks and the model evaluation measures that we employ are detailed in-depth in the third section, machine learning (3). Subsequent chapters on data and methods (4) cover the description of the NAPing PAnts system, preprocessing steps and collection of the data, and descriptions of the features and the target labels used in the supervised training process of the classifier. The experiments section (5) details the procedures used in the assessment of the impact of infant age on the feature data and the techniques used to mitigate this impact. In the results section (6), we disclose the numerical findings of this study, and the discussion (7) section contains the interpretations and reflection of these results and includes a discussion on the limitations and potential future research directions. Finally, a conclusion chapter (8) with the summary of findings finishes this thesis.

2 Background

2.1 Infant sleep

Sleep is a state of mental and physical engagement in which awareness is reduced. Throughout sleep, the majority of the body's systems are in a state of repair and growth, aiding in the restoration of the immune, nervous, skeletal, and muscular systems. These processes are crucial for maintaining mood, memory, and cognitive abilities and also have a significant impact on the functioning of the immune system.

Sleep occurs in repeating periods referred to as the sleep cycle. The sleep cycle is a recurring pattern of sleep stages that occur throughout the night. It consists of two main categories of sleep: Non-Rapid Eye Movement (NREM) sleep and Rapid Eye Movement (REM) sleep. NREM sleep is further divided into three stages: N1, N2, and N3. N1 is the transition from awake to sleep, characterized by light sleep and is easily disturbed by external stimuli. N2 is characterized by deeper relaxation and certain waveform characteristics detectable by electroencephalography (EEG). N3 is the deepest stage of NREM sleep, characterized by slow brain waves, and is essential for physical recovery and growth. Adults typically spend about 75% of their sleep in the NREM stages, with the majority in N2. After progressing through the non-rapid eye movement (NREM) stages, the sleep cycle transitions to rapid eye movement (REM) sleep, a nonrestful sleep stage characterized by vivid dreaming and temporary muscle paralysis that prevents physical movement during dreaming. The sleep cycle is not a linear process but repeats itself several times during a typical night's sleep. Each sleep cycle lasts on average 90 minutes in adults and 50 minutes in infants. The proportions of each stage within the sleep cycle change throughout the night. More time is spent in deep NREM sleep in the early part of the night, while more time is spent in REM sleep in the later part of the night. (Patel et al., 2022; Mervaala et al., 2019, pp. 220 – 223)

The sleep characteristics and architecture in newborns (0 - 2 months of age), infants (2 - 12 months) and toddlers (1 - 4 years) are highly dynamic and differ significantly from those of adults. Newborns do not have a 24-hour circadian rhythm like adults. Instead, they spend an equal amount of time sleeping during the day and night with an irregular rhythm, totaling approximately 16 to 18 hours per day. Infant sleep episodes typically last 3-4 hours, which is why newborns may wake up multiple times during the night. At this age, the sleep cycle consists of two stages: active sleep and quiet sleep. The cycle lasts an average of 50 minutes, compared to the 90-minute cycle of adults. Active sleep is similar to REM sleep in adults, while quiet sleep is similar to NREM sleep. Quiet sleep is characterized by stable respiratory frequency and lower variability in heart rate, while active sleep is characterized by variable respiration and heart rate and the presence of rapid eye movements. The distinct EEG activity associated with the different NREM sleep stages (N1, N2, and N3) does not become apparent until infants are two to three months old. This pattern gradually changes as newborns age and begin to develop a more adult-like circadian sleep rhythm. The amount of required sleep decreases while the length of sleep episodes increases and shifts toward a more nocturnal pattern. (MacLean et al., 2015, Patel et al., 2022, Tarullo et al., 2011)

Physiological factors, such as breathing and motor activity, also show age-related variations during the different phases of sleep in infants and newborns. According to a 2011 meta-analysis by Fleming et al. (2011), the respiratory rate undergoes a steep decline during the first two years, dropping from an average of 44 breaths to 26 breaths per minute. The overall breathing patterns during sleep also become more stable and apnoeic events where respiratory pauses occur become rarer (MacLean et al., 2015). Physical body movements and motor activity, such as body twitches and jerks, decrease (DeMasi et al., 2023).

2.2 Sleep measurement

Several techniques have been created to evaluate different aspects of sleep in infants and children. Polysomnography, actigraphy, keeping sleep diaries, and using questionnaires are widely recognized and used methods to assess sleep in this population. Assessing sleep in infants and children is a challenging task, and thus several techniques exist for different purposes in sleep assessment, each with advantages and limitations. (Sadeh, 2015)

Polysomnography is the gold standard for short-term sleep studies. During a PSG study, multiple physiological signals are typically recorded overnight in a sleep laboratory of a hospital. The recorded physiological signals include EEG for brain activity, electro-oculography (EOG) for eye movements, and electromyography (EMG) for muscle activity. In addition, respiratory factors, such as airflow and oxygen levels, are monitored, as well as cardiac activity, through an electrocardiogram (ECG). This data helps identify sleep stages and detect sleep disturbances such as sleep apnea. The identification of sleep stages mainly relies on EEG signal patterns, but also considers the other monitored signals. Sleep recordings are temporally divided into 30-second segments called epochs. Each epoch is classified into a specific sleep stage based on the recorded data by a trained medical expert. The current guidelines for this classification come from the American Academy of Sleep Medicine (AASM). (Mervaala et al., 2019, Sadeh, 2015)

Actigraphy relies on a wristwatch-like gadget that consistently tracks body movements and offers insights into sleep-wake cycles over long durations in the child's natural sleep environment. Advancements in technology have resulted in compact standalone devices that can conveniently be attached to the wrist or ankle in infants and young children to gather sleep activity information over extended periods with minimal disturbance. Actigraphy proves especially beneficial in evaluating sleep disorders due to its ability to maintain ongoing monitoring over prolonged periods. Traditional actigraphy is, however, limited by the fact that it does not provide data on sleep stages and therefore cannot provide crucial information on sleep quality and the sleep-wake cycle. (Sadeh, 2015)

Sleep diaries are used in long-term studies on sleep outside the hospital environment. Depending on the child's age, the diaries can be filled out by the child themselves or by the parent. Sleep diaries offer insights into the timing of sleep, nighttime awakenings, and other relevant aspects. The accuracy of parental entries in sleep diaries regarding their child's sleep routine, like bedtime and wake-up time, has been demonstrated to be reliable. However, the reliability of these journals decreases when it comes to assessing the quality of sleep. As questionnaires and diaries are not based on objective physiological measurements, but rather on the subjective judgments of parents, they do not provide an objective approach in long-term studies. (Sadeh, 2015)

2.3 Automatic sleep classification

Given the vast amount of data in a full-night sleep study carried out using PSG, for example, manual scoring and identification of different sleep stages can be laborious (960 30-second epochs in 8 hours of sleep). Thus, automated algorithms leveraging machine learning have been developed to aid the process.

In previous clinical studies, a variety of machine learning models have been successfully used, ranging from traditional algorithms such as the support vector machine (SVM) (Cortes and Vapnik, 1995) to deep learning methods, to study sleep-wake cycles and sleep stage classification.

For example, Aboalayon et al. (2016) studied sleep stage classification using EEG signals and proposed a system based on support vector machines that achieved a promising accuracy rate of 93% on 6 distinct stages of sleep. Werth et al. (2017) investigated the use of a non-linear kernel support vector machine to determine sleep stages (active and quiet sleep) based on known features of heart rate variability. Another study by Sekkal et al. (2022) examined the use of support vector machines compared to random forests and a long-short-term memory (LSTM) network, achieving an average precision of 81% to determine each stage (Wake, N1, N2, N3, REM) of sleep.

Previous research has shown that different stages of sleep are associated with distinct changes in several physiological variables, including respiration and body movement (Douglas et al., 1982, Haddad et al., 1987, Harper et al., 1987). Changes in these variables can be accurately measured and recorded using wearable sensors using an accelerometer or a gyroscope (Liu et al., 2011, Ranta et al., 2021a, Ryser et al., 2022). Machine learning-based sleep stage classification with portable sensor technology, such as wearable sensors and bed mattress-infused sensors, has previously been successfully studied in adult populations, paving the way for more objective long-term studies outside of the hospital setting (Mendez et al., 2010, Kortelainen et al., 2010, Tal et al., 2017, Gaiduk et al., 2018, Zhang et al., 2019).

Research on automatic sleep staging with less obtrusive physiological measurements that focus on tracking the sleep-wake cycle rather than all the different sleep stages has previously been for instance conducted by Kortelainen et al. (2010), Mendez et al. (2010), Tal et al. (2017) and Ranta et al. (2021b), each using a bed mattress sensor (BMS) utilizing a piezoelectric pressure-sensitive element. Kortelainen et al. (2010), for example, achieved an accuracy of 79% to distinguish REM sleep from NREM sleep based on features derived from the heart beat interval (HBI) and movement activity from the bed mattress sensor. Research by Tataraidze et al. (2015) and Yang et al. (2016) has demonstrated that sleep stage classification based solely on respiration derived features is feasible for adults. The classifier developed by Yang et al. used respiratory peak variance together with respiration for sleep stage classification, resulting in an overall accuracy of 71% in classification of sleep stages (Wake, NREM and REM). Tataraidze et al. achieved an initial classification accuracy of sleep stages (Wake, NREM, REM) of nearly 78% using a classifier with 33 respiratory-derived features.

Fewer studies in the field have been conducted on infants and young children, a population suffering more frequently from sleep disorders than adults. As noted earlier, the structure of sleep and the physiological indicators of different stages of sleep in infants, such as respiration rate, are dynamic and considerably different from those of adults, which introduce challenges for machine learning-based sleep stage classifiers (Tham et al., 2017, Baumert et al., 2023).

Focusing on monitoring sleep-wake patterns in infants in the neonatal intensive care unit (NICU), Ranta et al. (2021b) developed several classifiers based on SVM, a LSTM network, and a Convolutional Neural Network (CNN). Their version which relied solely on the motion data obtained from the bed mattress sensor was able to discriminate deep sleep (N3) from other sleep stages with an accuracy of 95%.

The NAPPA wearable was developed as a portable approach with automatic sleep staging for infants in out-of-hospital, long-term studies as its primary objective. A BiGRU-based sleep stage classifier associated with the system achieved a median accuracy of 77% in classifying stages of deep sleep (N2/N3), light sleep (N1/REM) and wake using five movement activity and respiration related features derived from the sensor signal. However, as detailed earlier in this chapter, age has a profound effect on sleep characteristics of infants and young children, which might cause machine learning models to misclassify sleep stages, especially in populations at the extremes of age. For instance, a rapid respiration in an older individual might be indicative of a certain sleep stage, but the same rate in a younger individual could indicate another (Schechtman and Harper, 1992, Litscher et al., 1993).

In conclusion, considerable advancements have been made in sleep stage classification using diverse machine learning techniques. However, the intricacies of physiological variables and sleep patterns, especially their strong dependence on age, emphasize that more research is needed on automatic sleep staging in the context of infants. By studying the integration age as a factor in sleep staging models, we can possibly strengthen the potential of wearable technology and automatic sleep staging for pediatric populations.

3 Machine learning

Machine learning is a branch of artificial intelligence that focuses on developing computational models based on algorithms that can learn from data via trainable parameters and make predictions or decisions. Traditional machine learning models, such as linear and logistic regression, random forests, and support vector machines are often used for various regression and classification tasks. These traditional models are usually classified under the umbrella of supervised learning, where the algorithm learns from a labeled dataset, understanding the relationship between the input features and the target output. (Goodfellow et al., 2016, pp. 98 – 101)

Supervised learning is a machine learning approach used for tasks where the dataset provided contains labeled examples. In this setting, each data point contains both features and a corresponding target value, the label. Features of a datapoint are typically a vector of predictor variables, whereas label is the outcome to be predicted. During training, the model learns to understand the relationship between the input features and their corresponding labels. Once trained, the model uses the learned relationship to predict the labels for new, unseen data, effectively applying the patterns it has learned from the training data to make accurate predictions. (Goodfellow et al., 2016, p. 105)

Supervised deep learning is a branch of machine learning that focuses on training artificial neural networks to automatically learn and represent complex patterns and features from data. These representations enable the model to capture intricate patterns and relationships, making deep learning particularly useful for tasks involving data with complex sequential or temporal structures and high dimensionality. (Goodfellow et al., 2016, p. 5) The "depth" in deep learning refers to the multiple layers of connected artificial neurons. This multilayered structure enables the network to derive hierarchical representations of the data, layer-by-layer. Initial layers might capture basic details, while deeper layers extract higher-level features. Deep learning is loosely inspired by neuroscience, and the idea of an artificial neural network that can learn relationships by the interactions of its computational subunits, neurons, was inspired by the brain. (Goodfellow et al., 2016, pp. 168 – 169)

While traditional machine learning models mostly rely on human-created features based on an existing understanding of the relationship between input and output data, deep learning models excel at learning representations and extracting features directly from raw data. This ability makes deep learning models generalizable across datasets, allowing them to perform effectively in scenarios where the relationship between input and output data is complex. As these models may contain thousands, millions or even trillions of parameters (at the writing of this thesis, GPT-4, the latest version of the popular generative AI platform ChatGPT, has been rumoured to contain over 1.7 Trillion parameters (Schreiner, 2023)), thus significant amounts of training data and computational resources are needed. The recent surge in the availability of computing power and data has made deep learning models popular across different fields of industry and research.

3.1 Neural networks

Neural networks represent a revolutionary step in the field of artificial intelligence, allowing for complex representations of data patterns through multiple layers of computation. The multilayer perceptron (MLP), which is the simplest class of deep learning neural networks, is showcased in Figure 1. The MLP, also known as a feedforward neural network, has a structure in which information passes from left to right through multiple fully connected layers. Each layer uses interconnected nodes, neurons, to process and transform input data in multiple stages towards the end of the network, with the output of one layer serving as the input for the next.

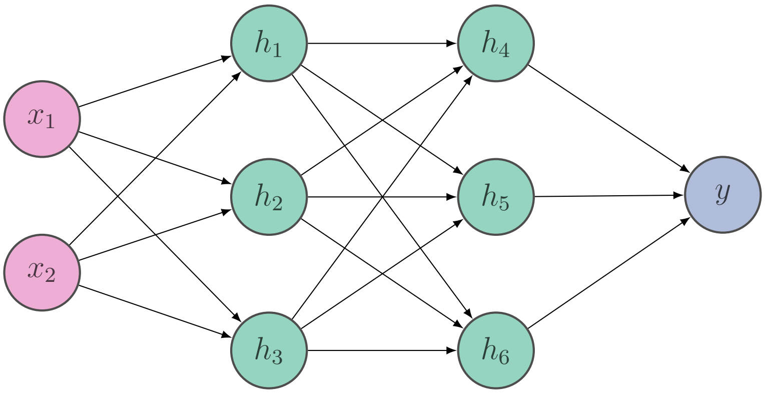


Figure 1: A simple feedforward artificial neural network consisting of four fully connected layers. The network has an input layer with two nodes (x_1, x_2), two hidden layers with six nodes and an output layer with one node (y). Such network could be for example trained to determine whether weather outside is good or bad given temperature and humidity as inputs. The number of nodes, layers and dimensionality of the input and output of the network is selected based on the application.

Each node of each hidden layer in a feedforward network contains learnable parameters, known as weights and the bias. The nodes may be thought of as independent functions operating in parallel, each taking in a d-dimensional input vector from the previous layer, transforming it via a linear combination of the weights and biases and an associated nonlinear 'activation' function ϕ

$$f(\mathbf{x}) = \phi(\sum_{i=1}^{d} w_i x_i + b) = \phi(\mathbf{w}^T\mathbf{x} + b) \qquad \mathbf{x}, \mathbf{w} \in \mathbb{R}^d. \tag{1}$$

In (1), $\mathbf{x}$ is the input vector of the node f, $\mathbf{w}$ are the weights of the node, b is the bias and ϕ is the activation function. The weights essentially determine the importance of a particular input in affecting the node's output. The bias, on the other hand, can shift the output up or down, influencing the node's final outcome. The purpose of the activation function is to preserve the the 'deep' layered structure of the network, as without the nonlinearity we would simply get a linear model.

The whole feedforward network $\mathbf{f}$:$\mathbb{R}^d \mapsto \mathbb{R}$ depicted in figure 1 may be algebraically expressed as a composition of its node functions:

$$\mathbf{f}(x;\theta) = f_4 \circ f_3 \circ f_2 \circ f_1 = f_4(f_3(f_2(f_1(x,\theta_1),\theta_2),\theta_3),\theta_4), \tag{2}$$

where the parameter θ consists of weights $\mathbf{w}$ and bias b. (Goodfellow et al., 2016, pp. 168 - 170)

Feedforward networks have been shown to be universal function approximators, and such network with finite number of neurons, layers and nonlinear activation functions can approximate any continuous function to an arbitrary degree of accuracy (Leshno et al., 1993).

3.2 Training neural networks

During the supervised training process, the objective is to optimize the parameters of the network so that its predictions closely match the true outcomes, or labels (Jung, 2022, p. 30). This process involves minimizing a measure of dissimilarity between the predicted outputs and the actual labels, known as the loss function. In the context of multiple class classification tasks, one commonly used approach for parameter estimation is maximum likelihood estimation, which aims to maximize the likelihood (3) of observing the given training data under the model's probability distribution (Goodfellow et al., 2016, p. 140). Due to challenges related to floating point arithmetic, it is more practical to minimize the negative logarithm of the likelihood instead, which converts multiplication to addition. This gives us the Negative Log-Likelihood Loss, also known as the Cross Entropy Loss (4) when used in conjunction with the softmax activation function (5). It is the most commonly used loss function in multiclass classification tasks.

$$\mathcal{L}(\mathcal{X};\theta) = \prod_{i=1}^{N} p_{model}(\mathcal{X};\theta) \tag{3}$$

$$\mathcal{L}(x,y;\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_c^{(i)} \log \sigma_c(\mathbf{f}(x^{(i)};\theta)) \tag{4}$$

$$\sigma(z)_c = \frac{\exp(z_c)}{\sum_{c'=1}^{C}\exp(z_{c'})}. \tag{5}$$

In the formulae above, N is the number of training examples in the training set:

$$\mathcal{X} = \{(x^{(1)},y^{(1)}),\ (x^{(2)},y^{(2)}),\ \ldots,\ (x^{(N)},y^{(N)})\},$$

where C denotes the total number of classes in the classification task. Vector $y^{(i)}$ is the one-hot encoded i:th target and $x^{(i)}$ is the d-dimensional input vector to the network. The one-hot encoding means that every element in the vector is zero except for the index of the class in the i:th example in the set, which is one.

The neural network $\mathbf{f} : \mathbb{R}^d \mapsto \mathbb{R}^C$ takes the input vector and maps it to a logit output vector of the same dimension as the number of classes. In the softmax function, z represents the vector of logits or raw scores produced by the network before applying any activation function. Each element z_c corresponds to the raw score for class c. The softmax function exponentiates these scores and normalizes them by the sum of exponentiated scores over all classes. This normalization ensures that the output values of the softmax function sum up to 1 and lie in the range [0, 1], thus representing a confidence level over the classes in the classification task (Goodfellow et al., 2016, p. 184). In essence, the softmax function converts the logits into confidence values, indicating the likelihood of each class for a given input x.

Finding the parameters for the network that minimize the loss function requires solving the following optimization problem

$$\theta^* = \text{argmin}.\ \mathcal{L}(\theta).$$

simple method of solving the optimization problem at hand is to use the gradient descent method:

$$\theta_{t+1} = \theta_t - \epsilon \nabla_\theta \mathcal{L}(\theta_t),$$

which updates the network's parameters by taking a step in the direction where the value of the loss function $\mathcal{L}(\theta)$ decreases the fastest. The step size is multiplied by the learning rate ϵ, which greatly influences the convergence of the algorithm towards the search of the optimum.

Because modern training datasets may contain millions or billions of data points, computing the gradient for the whole data is not practical due to computational memory limitations. Instead, the gradient of the loss function is approximated by using a small subset of the original training dataset called a minibatch, for which data points are randomly selected. The model parameters are updated with each new minibatch, until all data points in the training set have been used once. This known as an epoch of training. Multiple epochs are used to ensure that the model has enough iterations and updates to effectively minimize the loss function. This method of using mini-batches in training introduces noise into the gradient computations, which may help the model to generalize better outside the training data. This technique is known as stochastic gradient descent (SGD). (Goodfellow et al., 2016)

Adam (Kingma and Ba, 2014) is a popular optimizer based on SGD that is used in the training of deep neural networks. It introduces adaptive moment estimation by maintaining exponentially decaying moving averages of the first moment and the second moment of the gradients. This approach allows Adam to dynamically adjust learning rates for different parameters during training, making it a robust optimizer. (Goodfellow et al., 2016, p. 309)

Backpropagation (Linnainmaa, 1970, Rumelhart et al., 1986) is an algorithm used to calculate the loss function gradients by propagating the error from the output layer back to the input layer. It calculates how much each parameter contributes to the overall error and then updates the parameters in the opposite direction of the gradient to minimize the loss. This process is iteratively repeated for each data point in the training set until the parameters converge to values that minimize the loss function and make the model perform well on the given task. Backpropagation is based on the chain rule of differentiation.

3.3 Recurrent neural networks

Recurrent Neural Networks (RNNs) are a class of deep learning neural networks designed to process sequential data by incorporating feedback connections that allow information to persist over time. Unlike traditional feedforward neural networks (akin to the network in Figure 1), where data flows only in one direction from input to output, RNNs process sequences one element at a time while maintaining a hidden state, which acts as the memory of the network, that encodes information about past elements in the sequence. RNNs employ parameter sharing across different time steps, which means that the same set of parameters (weights and biases) is used for each element in the input sequence. This feature allows RNNs to effectively handle sequences of varying lengths, and the shared parameters also help in reducing the model's complexity and the amount of data required for training. The ability of 'remembering' information while processing sequences of inputs makes RNN models suitable for tasks in medical time series analysis, such as sleep stage classification, where the order of the input or the context of each data point in the sequence carries important information. (Goodfellow et al., 2016, pp. 397 – 400)

At the core of an RNN is the recurrent connection, shown in Figure 2, which enables the network to process sequential data of varying lengths and to capture temporal dependencies between elements. The recurrent connection introduces a feedback loop that allows the output of the previous time step to influence the computation of the current time step. This feedback mechanism allows RNNs to retain information about earlier elements of the sequence at each time step in the fixed-size 'hidden state' vector, often denoted as h_t.

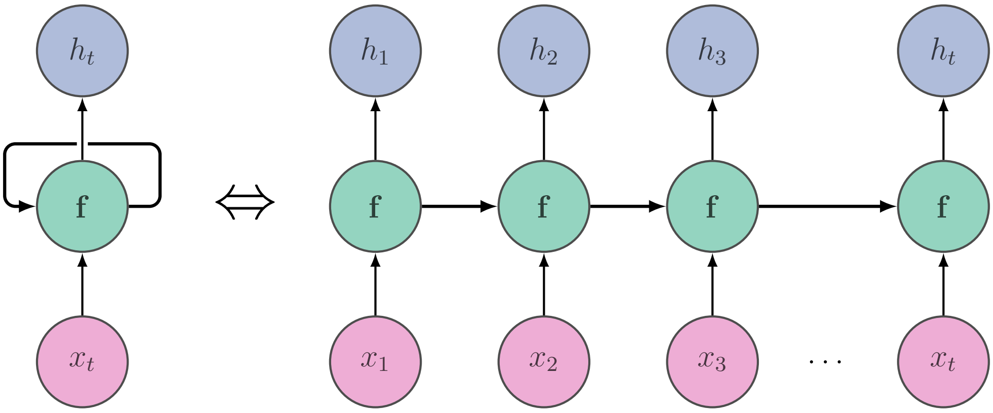


Figure 2: Left: Graphical representation of a Recurrent Neural Network illustrating the feedback connection. Right: Unrolled RNN showing the hidden states (h_t) corresponding to the network input data (x_t) at time t.

Given a sequence of input data (x_1, x_2, ..., x_t), the RNN processes each element sequentially and updates the hidden state at each time step. The hidden state (h_t) at the final time step t captures essential information about the entire input sequence, serving as a compact representation relevant to the learning task. The size of the hidden vector is a key factor that influences the network's ability to process and remember information over time, as it determines capacity of the network's 'memory' at each time step.

The standard RNN type of design, however, only makes use of the past elements in the sequence. In many real-world scenarios, the prediction or understanding of a particular element in a sequence can be heavily influenced not only by the preceding elements, but also by the following elements.

Bidirectional RNNs (BiRNNs) extend the standard RNN architecture by processing sequences in both directions by using two recurrent neural networks simultaneously. At each time step, a BiRNN integrates information from both the forward and backward directions and produces the final output by concatenating the outputs of the two RNNs. This two-way processing as shown in Figure 3 offers enhanced context and enables the model to better capture the sequential or temporal dependencies in the data. The trade-off in bidirectional models is increased model complexity and longer training times. (Goodfellow et al., 2016, p. 383)

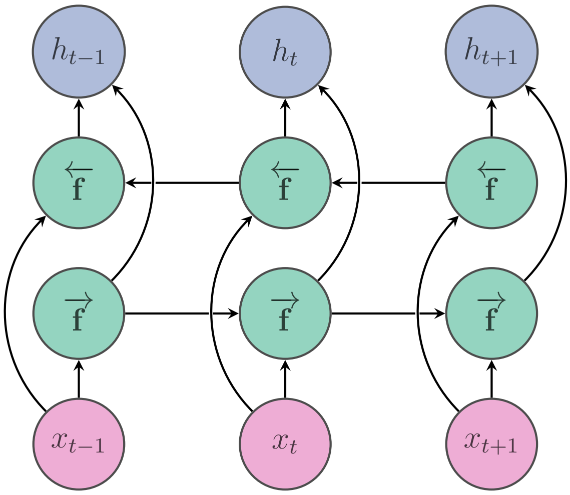

h_{t-1}
h_t
h_{t+1}
$\overleftarrow{\mathbf{f}}$
$\overleftarrow{\mathbf{f}}$
$\overleftarrow{\mathbf{f}}$
$\overrightarrow{\mathbf{f}}$
$\overrightarrow{\mathbf{f}}$
$\overrightarrow{\mathbf{f}}$
x_{t-1}
x_t
x_{t+1}

Figure 3: A graphical representation of a bidirectional RNN. In this network, the hidden states are computed by the network proceeding forwards and backwards in the input data simultaneously by using two neural networks $\overleftarrow{\mathbf{f}}$ and $\overrightarrow{\mathbf{f}}$.

Adjusting the parameters of RNNs during the training process can be done in a similar fashion as with the traditional feedforward networks by using the stochastic gradient descent and its variants. The difference due to the recurrent nature of RNNs, where each layer repeats the same function with shared weights, is that the backpropagation is applied through time. This approach can be thought of as unrolling the RNN across time steps and applying gradient descent as if it were a deep feedforward network. This recurrence, however, introduces a challenge when the lengths of the input sequences to the RNN grow, as the effect of applying the shared weights to the input over and over again accumulates and can cause the values of the output of the network to grow or decay exponentially while it proceeds further towards the end of the sequence. This can result in a problem known as the vanishing or exploding gradient, which causes gradients to shrink or grow uncontrollably as they are propagated back through time, resulting in the stochastic gradient descent based optimizer to take either too big or small steps in search of the optimum, thus hindering convergence and therefore severely disrupting the network's ability to learn long-range dependencies. Another drawback with RNNs is the fact that the computations have to be carried out sequentially. This nature of RNNs restricts the extent to which parallel computing, a feature that significantly accelerates training and inference in traditional neural networks, can be leveraged. (Goodfellow et al., 2016)

The field of deep learning has introduced several techiniques and neural network architectures to mitigate the exploding or vanishing gradient problem in RNNs. Methods such as gradient clipping, where a gradient vector exceeding a certain magnitude is clipped prior to updating the model parameters, have been found successful. The Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) network, which uses specialized memory cells and gating mechanisms that selectively control the flow of information to capture long-term dependencies was a significant advancement for sequential data-analysis tasks such as machine translation and speech recognition. The LSTM cell maintains two states, the hidden state and the cell state, which act as it's 'memory'.

3.4 Gated recurrent unit

The Gated Recurrent Unit (GRU), illustrated in Figure 4, was introduced by Cho et al. (2014) to address issues in traditional RNNs such as the problem of vanishing and exploding gradients. GRUs, like LSTMs, are designed to effectively capture long-term dependencies but with a simpler architecture. Instead of maintaining both a cell state and a hidden state, the GRU cell stores all of the information from previous states in the hidden state (likewise to the general RNN principle introduced previously). This allows reducing the number of parameters and neural network units within the cell, making the computations during training less complex and the model potentially more robust against overfitting in scenarios with limited training data. Some research has suggested that the GRU outperforms LSTM in certain tasks where datasets are smaller in size and input sequences are shorter and less complex (Yang et al., 2020, Cahuantzi et al., 2023).

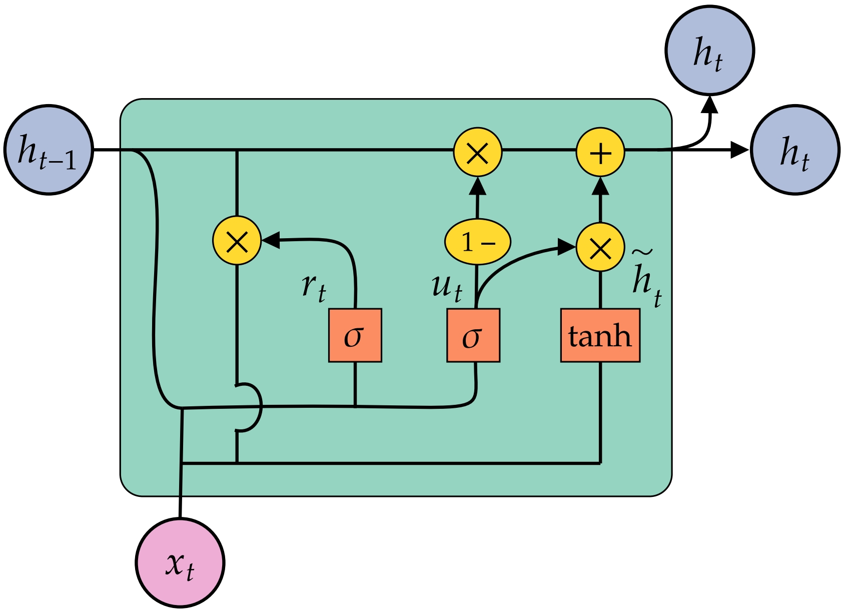


Figure 4: The recurrent unit of the GRU network, with orange elements representing neural network layers using nonlinear activation functions (the sigmoid and the hyperbolic tangent) and yellow elements indicating pointwise operators.

A GRU cell containts two primary gates: the reset gate (r_t) and the update gate (u_t). The reset gate serves as the controller for the hidden state's memory, and it determines which parts of the previous hidden state should be forgotten or "reset" based on the current input and the previous hidden state. The update gate regulates the extent to which the current input should influence the new hidden state. It decides which information from the input should be integrated into the hidden state, effectively controlling the memory update process. This gating logic is useful in scenarios with multiple processes occurring at the same time and at varying rates, allowing selective updates to the hidden state to reflect rapid changes while maintaining stability for slower evolving processes. Each gate uses a nonlinear activation function, which is either the sigmoid, $\sigma(\text{x})$, or the hyperbolic tangent, $\tanh(\text{x})$, which are defined as:

$$\begin{aligned} \sigma(x) &= \frac{1}{1+e^{-x}} \\ \tanh(x) &= \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}. \end{aligned}$$

The update rules for the update gate and the reset gate are defined as follows:

$$\begin{aligned} r_t &= \sigma(\mathbf{W}_r h_{t-1} + \mathbf{U}_r x_t + \mathbf{b}_r) \\ u_t &= \sigma(\mathbf{W}_u h_{t-1} + \mathbf{U}_u x_t + \mathbf{b}_u), \end{aligned}$$

where the elements of the matrices $\mathbf{W}$ and $\mathbf{U}$ and the bias vectors $\mathbf{b}$ are learnt during training. The final hidden state, h_t, is computed as a combination of the previous hidden state (h_{t-1}) and the new candidate state ($\tilde{h}_t$), which is influenced by both the reset gate and the current input, with the update gate (u_t) determining the balance. The new hidden state is given by:

$$\begin{aligned} \tilde{h}_t &= \tanh(\mathbf{W}(r_t \odot h_{t-1}) + \mathbf{U}x_t + \mathbf{b}_h) \\ h_t &= (1-u_t) \odot h_{t-1} + u_t \odot \tilde{h}_t, \end{aligned}$$

where $\odot$ is the Hadamard (elementwise) tensor product.

The GRU can be utilized in a bidirectional manner likewise to the general BiRNN principle introduced in chapter 3.3, Figure 3. Additionally, multiple GRU units can be stacked one on top of another to form a model with a deep layered structure, where the output of one GRU layer serves as the input for the next layer and so on. The use of layered structure further increases the capacity to learn complex patterns in data, as each layer can capture different levels of abstraction.

3.5 Performance metrics

The performance of machine learning models is evaluated using performance metrics. In classification tasks, these metrics provide valuable insight into the model's ability to correctly classify instances from different classes. We use the following criteria to evaluate and compare the models in our study:

Accuracy represents the proportion of correctly classified instances out of the total number of instances in the dataset. It provides a simple and intuitive understanding of the model's overall performance; however, it is not a reliable measure of performance for imbalanced datasets and should not be used as the sole measure of performance. This is because a given model can obtain a good accuracy score by performing well on classifying the instances of the majority class, while neglecting the minority classes. This imbalance issue is prevalent in general in medical datasets and particularly in automatic sleep stage classification, as the distribution of different sleep stages can be highly uneven. This is due to the fact that most of the night's sleep is typically spent in NREM sleep (Patel et al., 2022), leading to a skewed dataset where the majority class dominates. In such cases, relying on accuracy alone could give an overly optimistic picture of the model's performance. Accuracy is calculated as

$$\text{Accuracy} = \frac{\#\text{ Correctly classified instances}}{\#\text{ All instances}}.$$

Matthews Correlation Coefficient (MCC) (Matthews, 1975) is a robust performance metric that takes into account class imbalance, providing a good and balanced measure of performance even in scenarios of highly imbalanced datasets. The MCC ranges from -1 to 1, where 1 indicates a perfect prediction, 0 indicates a random prediction, and -1 indicates a complete disagreement between the model's predictions and the true labels. In the binary case, MCC is equivalent to the Pearson correlation coefficient estimated for two binary variables. The MCC is computed with the following simple formula:

$$\text{MCC} = \frac{\text{TN} \times \text{TP} - \text{FN} \times \text{FP}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}},$$

where TN indicates the number of true negative predictions, TP true positives, FP false positives and FN false negative predictions. The MCC metric can be extended for multiclass classification problems via:

$$\text{MCC} = \frac{c \times s - \sum_k^K p_k \times t_k}{\sqrt{(s^2 - \sum_k^K p_k^2) \times (s^2 - \sum_k^K t_k^2)}}$$

where $t_k = \sum_i^K C_{ik}$ the number of times class k occurred, $p_k = \sum_i^K C_{ki}$ the number of times it was predicted, $c = \sum_k^K C_{kk}$ the number of samples correctly predicted and $s = \sum_i^K \sum_j^K C_{ij}$ is the total number of samples. (Pedregosa et al., 2021)

Confusion Matrix is a visualization tool that can be used to assess the performance of a classification model. Table 1 represents a confusion matrix for a classification problem with three classes: A, B and C. Each column of the matrix represents the instances in the actual class, while each row represents the instances in the predicted class, with the diagonal elements representing the agreement between the true and predicted values (T) while off-diagonal elements indicate false predictions (F). This matrix provides insight into the errors made by the classifier model and is particularly useful in assessing the types of misclassifications that the model tends to make.

Class	**A**	**B**	**C**
A	T	F	F
B	F	T	F
C	F	F	T

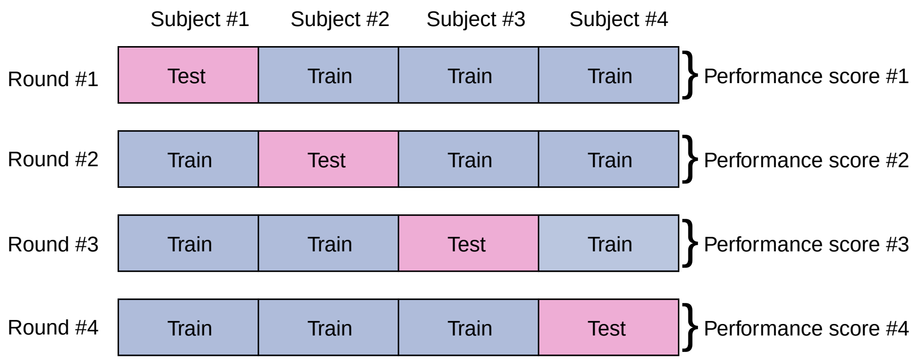
Table 1: 3x3 Confusion matrix for a multiclass classification task. The total number of rows and columns of the matrix correspond to the number of classes in the classification task. Values of the elements are typically the column-wise normalized percentages, the number of instances or both.

3.6 Leave-one-subject-out cross-validation

Training and evaluating the performance of a machine learning model requires that part of the available data is used to train the model, while another part of the data is used to measure the performance of the trained model. These sets must not overlap, as the evaluation should be performed on 'unseen' data by the model. This is because machine learning models can start to learn and model the statistical noise or idiosyncrasies in the data, while failing to learn the true meaningful relationships between the features and the labels, thus resulting in the model not being able to generalize outside the dataset it was trained on, resulting in overfitting of the model.

Cross-validation can be used to perform model training and subsequent evaluation. In standard k-fold cross-validation, the available data is divided into k subsets or "folds" of approximately equal size. The model is trained and evaluated k times, with each fold serving as a test set once and the remaining $k-1$ folds are used for training. As a result, cross-validation yields k number of trained models. Finally, the overall generalization performance of the model can be assessed by using each generated model to make predictions on its corresponding test fold, the data of which was not used train the model. The predictions made by the model are compared to the true labels with respect to some suitable performance metric, such as the MCC or accuracy. Finally, an overall estimate of model performance can be measured by computing population statistics across the fold metric results, such as the mean and the standard deviation. (Jung, 2022, pp. 159 – 160)

Leave-One-Subject-Out Cross-Validation (LOSOCV, visualized in Figure 5) is a special case of k-fold cross-validation, where k is set equal to the number of subjects in the entire dataset. For each iteration, the data of a single subject is retained as the test set, while the rest of the subjects' data is used for training. This process is repeated for all subjects in the dataset. The LOSOCV evaluation paradigm is characterized by low bias, as in each iteration the size of the test set is minimized, while size of the training set is maximal. However, this approach is computationally expensive and can result in a high variance in performance due to individual differences. Despite these challenges, LOSOCV remains a preferred method when the dataset size is limited.



	Subject #1	Subject #2	Subject #3	Subject #4	
Round #1	Test	Train	Train	Train	} Performance score #1
Round #2	Train	Test	Train	Train	} Performance score #2
Round #3	Train	Train	Test	Train	} Performance score #3
Round #4	Train	Train	Train	Test	} Performance score #4

Figure 5: An illustration of LOSOCV procedure with data from four distinct subjects.

4 Data and methods

4.1 NAPPA wearable

The Napping Pants NAPPA wearable is seen in Figure 6. The device is specifically designed for infants and is used for minimally disruptive and easy recording of respiration and movement activity during sleep. The wearable is equipped with a detachable, battery operated Movesense movement sensor (manufactured by Movesense Ltd.), which is certified for health monitoring purposes. The sensor is held in place by a textile garment, which covers the diapers. The placement and sensor type allow the tracking of abdominal respiratory movements with a high accuracy, offering a way of monitoring the infant during sleep in extended periods in a non-clinical setting.

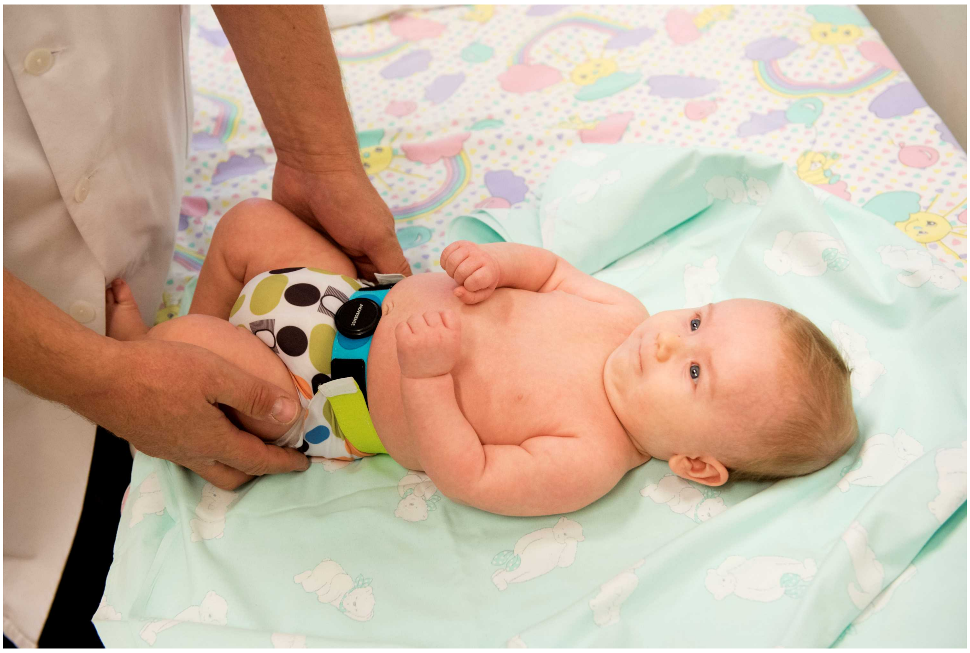


Figure 6: An infant wearing the NAPPA system. The movement sensor is attached to the front of the diaper. Picture: https://www.babacenter.fi

The sensor uses both a gyroscope and an accelerometer for capturing triaxial movement data. Specifically, it sensor records linear acceleration and angular velocity across the x, y, and z axes with an adjustable sampling rate. The sensor is programmable, has Bluetooth for wireless communication and internal data storage. The recorded movement data are stored either on a mobile device via a Bluetooth connection or directly logged into the sensor memory, from which the data can be retrieved afterwards.

From the raw sensor signal, sleep relevant features such as movement activity and respiration rate can be computed either by firmware embedded in the sensor or by a Matlab-based algorithm afterwards. The features are calculated in 30 second epochs to match the scoring interval of sleep stages in PSG studies, thus allowing the labeling of each recorded data point with a corresponding sleep stage.

4.2 Data collection and preprocessing

The clinical data used in this study was collected during the years 2019-2023 in the sleep laboratory of the Department of Clinical Neurophysiology at New Children's Hospital, part of the Helsinki University Central Hospital. Data from a cohort of 36 distinct infants recruited to a polysomnography study, ranging in age from 2 weeks to 20 months (median age 3 months), was used for classifier training and subsequent evaluation. The distribution of subject ages and sleep recording lengths is visualized by boxplots in Figure 7.

Following established clinical procedures, PSG data were recorded at a 200 Hz sampling rate using the Embla N700 system coupled with the RemLogic 3.2.0 software from Natus. Sleep stages, including wake, N1, N2, N3, and REM, were identified and labeled in 30-second epochs by a clinical expert according to the guidelines set by the American Academy of Sleep Medicine (AASM). The scoring procedure yields a *hypnogram*, a data file that represents the stages of sleep as a function of time. The PSG studies were carried out while the infants were wearing the NAPPA portable system to allow the labeling of each recorded data point corresponding to the sleep stages determined by the clinician, allowing supervised training of the classifier model.

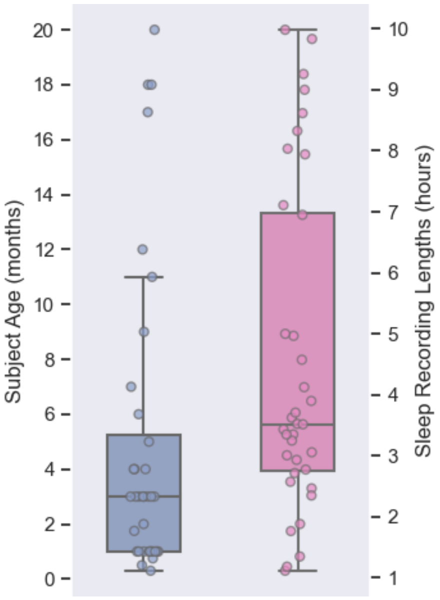


Figure 7: Box plots depicting the distribution of infant ages and sleep recording lengths.

The signals from all sleep recordings were visually inspected for manual artifact removal. For example, in an instance where the recording of NAPPA sensor was halted for a period of an hour during one clinical study due to a malfunction, we cut out this period with missing data. Furthermore, as the feature computation or the sampling rate of the Movesense sensor turned out not to be consistent, leading to epoch lengths deviating from the 30 second standard by as much as ± 1 second, the sensor data had to be resampled to match the scoring interval of the *exact* 30 seconds in the hypnograms for temporal alignment. We used Pandas (Mckinney, 2011), a Python library widely used in data analysis and machine learning, to convert the sensor data frequency to be compatible with the frequency of the hypnograms. In cases where downsampling was necessary (sensor sampling frequency greater than hypnogram frequency), we used rolling mean. Cases where the sensor sampling frequency was lower than the hypnogram frequency, we used nearest neighbor interpolation.

In total, these NAPPA sleep recordings yielded a dataset containing 19 783 labelled data points of the sensor data. The lengths of the sleep recordings ranged from 1 hour to 10 hours, with a median recording length of 4 hours.

4.3 Features

The input features for the sleep classifier are calculated from the raw signal of the Movesense movement sensor, which has a sample rate set at 13 hz. Ranta et al. (2021a). selected a total of five distinct body movement and respiration behavior describing features by using the set of features previously engineered and extracted for a bed mattress sensor-based sleep stage classifier (Ranta et al., 2021b). The feature selection process considered a large collection of features that were physiologically related, for example, to respiration, heart rate, and movement, and considered both temporal and spectral (frequency/Fourier) domains. The final features were selected using the Minimum Redundancy Maximum Relevance (MRMR) principle.

Of these selected five features, four originate from the y-axis channel of the gyroscope and one from the triaxial accelerometer signal. The accelerometer-derived movement activity feature assesses the overall abdominal movement activity of the infant during sleep, while the gyroscope-derived features are related to the stability and regularity of respiration, which are helpful in determining different stages of sleep. The y-axis of the gyroscope, oriented towards the head of the infant and measuring the pitch angular velocity, has been identified as the most representative of respiration movements, making it a primary source for the feature set (Acosta-Leinonen, 2019).

The first feature, **Movement Activity** is computed by taking the magnitude of the triaxial accelerometer sample at a given time t:

$$\|a_t\| = \sqrt{x_t^2 + y_t^2 + z_t^2}. \tag{6}$$

To extract the relevant frequencies, 1-6 Hz Butterworth band-pass filter is applied and the resulting magnitude signal is rectified. To get the movement activity in units m/s, the values are integrated in a moving 5 second window:

$$\text{activity}_t = \sum_{T \in t \pm 2.5s} \|a_t\| \frac{1}{f_s},$$

where f_s is the sample rate of 13 Hz. Finally, the activity values are averaged over 30 second windows to match the epoch lengths in the PSG studies.

The second feature, **Respiration Autocorrelation Function Maximum** originates from the y-axis of the gyroscope, and is computed from 30 second non-overlapping windows. Prior to feature calculation, the signal is band-pass filtered with second-order Butterworth in frequency range of 0.1 - 1.5 Hz. The autocorrelation of a sample at lag k is given by the autocorrelation function:

$$ACF(y_t, y_{t+k}) = \frac{c_k}{c_0} = \frac{\frac{1}{T-1}\sum_{t=1}^{T-k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y_t})},$$

where c_0 is the sample variance and c_k the covariance of the time series at lag k. The correlations are calculated up to 4 seconds, and therefore lags $k \in [1, f_s \times 4s]$ are considered. T is equal to $f_s \times 30$ s, as each data point describes a time window of 30 seconds. The autocorrelation is computed by Matlab function 'autocorr()' and the function 'max()' is used to obtain the maximum value, m.

Autocorrelation measures the correlation of time series with a time-shifted version of itself, i.e. the similarity between observations as a function of the time lag between them. The intuition behind this feature is that a high autocorrelation value indicates a regular, more periodic breathing whereas a low value is indicative of irregular breathing patterns.

The third feature is **Respiration Rate**, and it is computed using the corresponding time lag of the maximum autocorrelation value (m) for each 30 second non-overlapping window. The frequency is multiplied by 60 seconds to get the value in units breaths per minute:

$$f_{RR} = \frac{1}{m/f_s} \times 60s$$

The fourth and the fifth features are **Respiration Peaks Median** and **Respiration Peaks Standard Deviation**, which also originate from the y-channel of the gyroscope. The signal is filtered with the second order Butterworth in a frequency range of 0.1 - 1.5 Hz, and the peak values of the filtered signals are found with the Matlab function 'findpeaks()'. From the array of peaks, median and standard deviation values are computed. These features are also related to the stability and regularity of breathing, where high median peak values indicate sharper belly movements and a high standard deviation indicates greater variability in these movements.

The logarithmic distribution of the feature data is visualized by histograms in Figure 8.

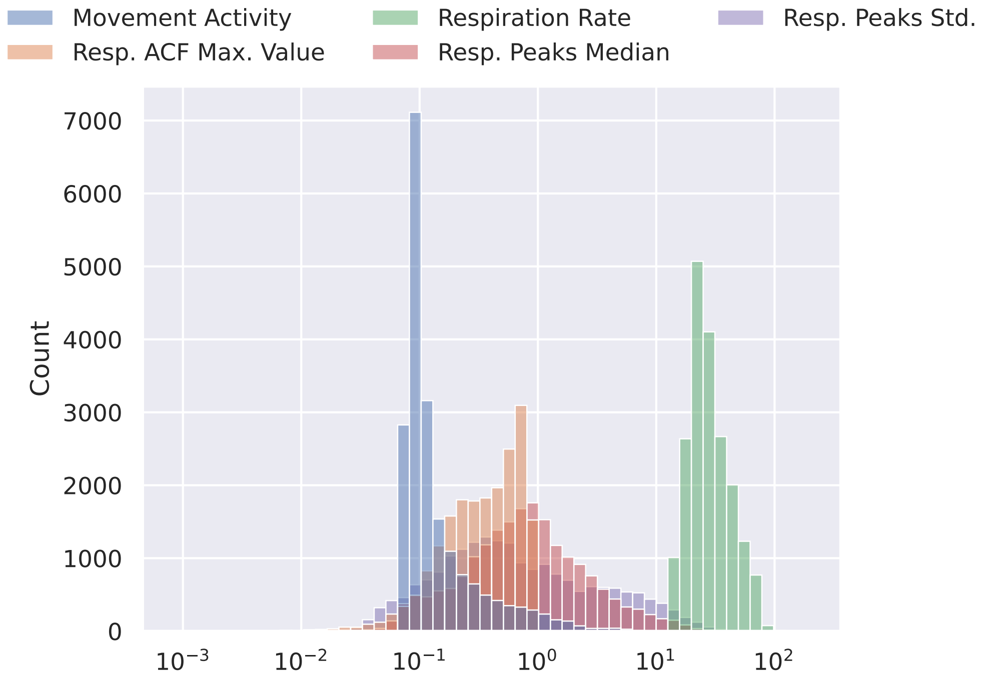


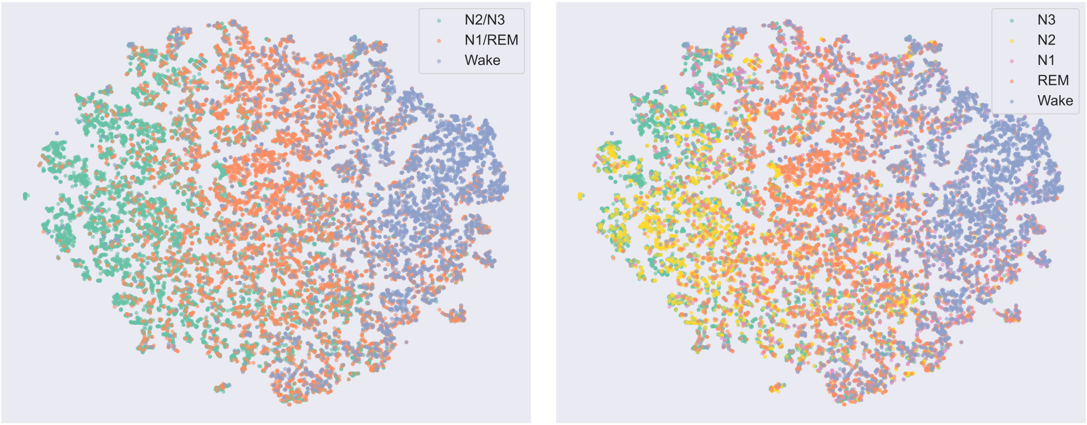
Figure 8: Multivariate histogram depicting the logarithmic distribution of features.

4.4 Labels

Instead of attempting to classify all of the five individual sleep stages standardized by the AASM, the NAPPA system focuses on distinguishing between three unique states of alertness: deep sleep (N2/N3), light sleep (N1/REM), and wake.

The reasoning for categorizing sleep into these particular three classes has several justifications: First, a key component of assessing sleep quality is the ability to distinguish the amount of deep sleep compared to periods of wakefulness or light sleep. Deep sleep plays a crucial role in restoring and rejuvenating the body, and therefore, the amount of deep sleep provides valuable prognostic information and insight into an individual's overall health. Prolonged periods of wakefulness or light sleep on the contrary may indicate potential underlying health problems. (Patel et al., 2022). Second, the inherent limitations of the movement-based features influenced the classification strategy. A nuanced classification that included all the different sleep states would require access to a broader range of physiological information. For example, distinguishing N1 from REM sleep requires information on eye movement, while separating N2 from N3 sleep depends on cortical EEG activity data (Patel et al., 2022, Mervaala et al., 2019).

This reasoning in selecting the target labels is further supported by a t-SNE (Hinton and Roweis, 2002) visualization of our data set (Figure 9), which projects the five-dimensional feature space down to two dimensions. We observe that the separability between states is greater with merged sleep states (Figure 9a) compared to the full five sleep states (Figure 9b), where a high degree of class overlap is visible.



(a) 3-state sleep labels

(b) 5-state sleep labels

Figure 9: t-SNE visualization of the highdimensional features.

With combined sleep classes as target labels for the classifier, the NAPPA system aims to strike a reasonable balance between classification accuracy and the ability to assess overall sleep health in infant populations.

Figure 10 visualizes the classifier target labels along with the input features during a sleep recording lasting approximately 8 hours. Distinct changes in the features are apparent with different sleep stages. During stages of deep sleep (N2/N3), higher autocorrelation values are present, while movement activity and respiration rate are stable. During periods of wakefulness (at 03:30 am and 05:00 am), movement activity values peak, while autocorrelation values drop, indicating a nonregular breathing pattern. Respiration rate fluctuates and the respiration peak median and standard deviation values increase.

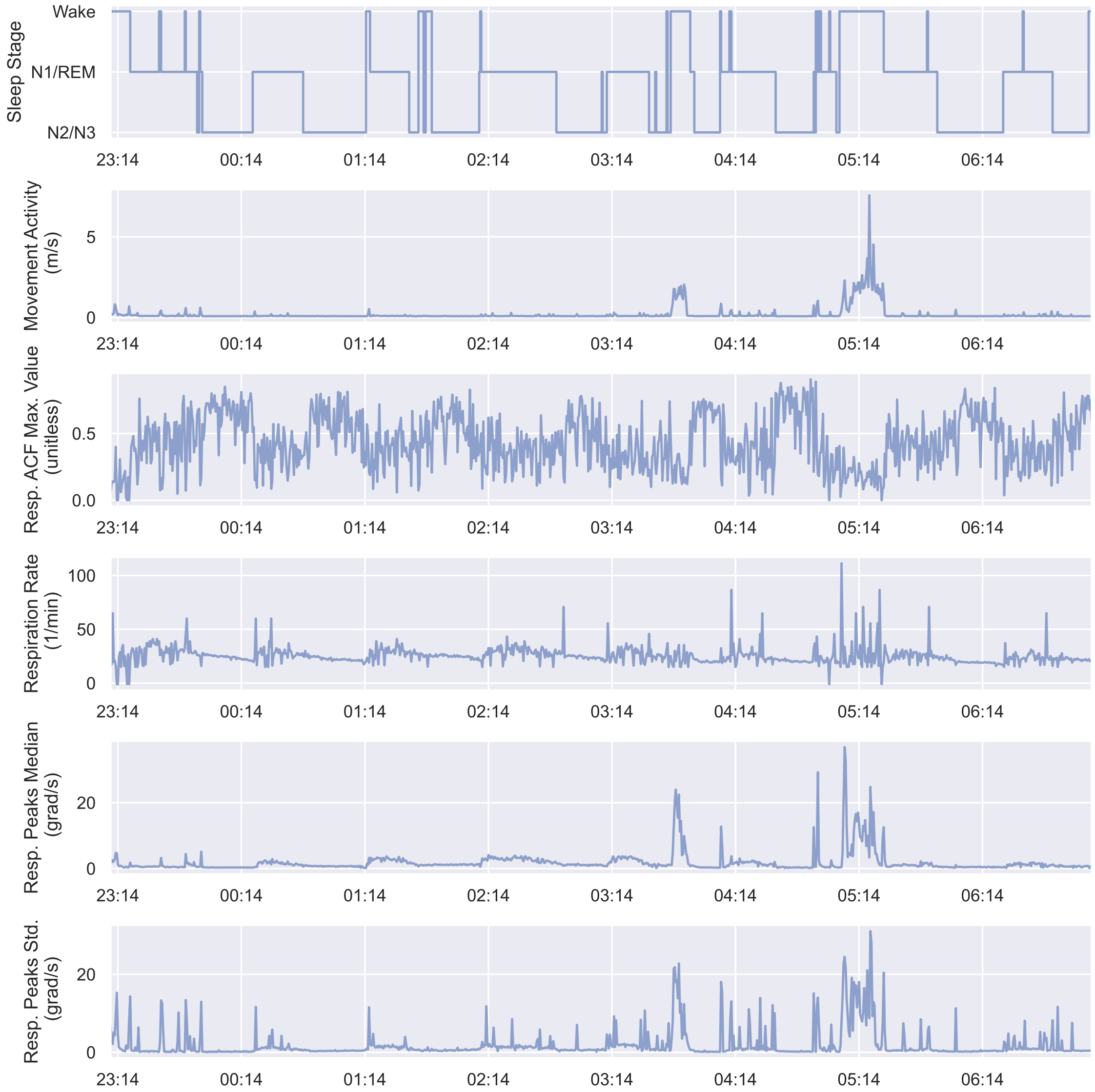


Figure 10: PSG annotated sleep stages along with feature data from the NAPPA wearable depicted as time series. The topmost plot visualizes the (merged) sleep stages and transitions between, while the other plots visualize the feature data during an 8-hour night's sleep.

4.5 Sleep stage classifier

We have selected a BiGRU with two layers as the primary architecture for the sleep classifier model, using the five distinct sleep features detailed in section 4.3 for identifying one of the three merged sleep stages at a given time step in each of the sleep recordings.

A recurrent neural network, like the GRU, is an appropriate choice because transitions of sleep stages during the progression of sleep are temporally dependent. We leverage the BiRNN principle introduced in section 3.3 to offer more context for the classifier in each sleep sequence when making predictions. To enhance the model's capacity to potentially learn more intricate patterns in the input data, we utilize a layered design with two such BiGRU units, where the input of the second BiGRU unit is the output of the first BiGRU unit.

At the final stage of our model, a fully connected linear output layer is used to classify the sleep stages by transforming the hidden state h_t output of the final BiGRU layer at every time step in each individual sleep sequence. This final layer produces a three-dimensional vector for each time step, with each dimension representing a logit corresponding to one of the sleep stages: wake, light sleep (N1/REM) or deep sleep (N2/N3). Finally, the output logits of the fully connected layer are passed through a softmax activation function to generate a probability distribution over the sleep classes. The class with the highest probability is considered the final prediction made by the model.

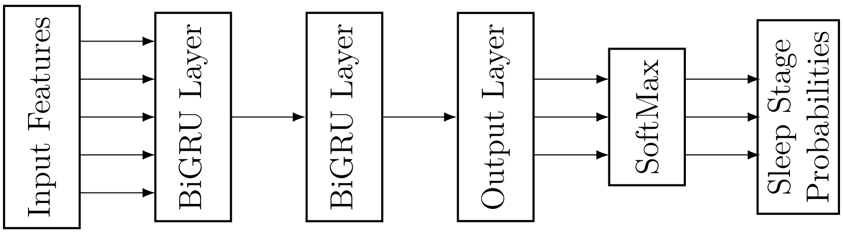


Figure 11: A simple illustration of the architecture of the sleep stage classifier model.

5 Experiments

5.1 Evaluating age influence on features

5.1.1 Correlation analysis

As noted in chapter 2, previous research has shown that the physiological behavior and indicators of different sleep stages in infants and children vary across different age groups. In this section, we analyze the relationship between subject age and features in our sleep recordings. To quantify these relationships, we employ the Spearman correlation coefficient (7), a nonparametric statistic denoted by ρ. In Spearman correlation, before calculating the coefficient, each data item is replaced by its rank. The ranks are the position indexes of each data item when the data is sorted. Spearman correlation is given by the formula:

$$\rho = \frac{\mathrm{COV}(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}, \tag{7}$$

where $R(X)$ and $R(Y)$ are the rank values of X and Y, respectively. The value of ρ is in range $[-1, 1]$ and evaluates the strength and direction of the monotonic relationship between two ranked variables.

We chose Spearman's correlation over the traditional Pearson correlation, because the Spearman correlation is more robust against nonlinear relationships and outliers in the data. In addition, Spearman's correlation is nonparametric, meaning that it does not rely on assumptions about the underlying distribution of the data, whereas Pearson's correlation requires both datasets to be normally distributed. Therefore, the Spearman correlation provides a more appropriate measure of association in cases where these assumptions cannot be confirmed to hold (Mohr et al., 2022, pp. 670 – 672).

To assess the statistical significance of the correlations, we used the permutation test, which is a nonparametric method suitable for small sample sizes. In the permutation test, the correlation coefficient is computed from the original data sets to obtain the observed value. Then the rank values of one of the variables is shuffled n times, each time computing the correlation coefficient. The instances where the correlation is at least as extreme as the original observed correlation (has an absolute value equal to or greater than the original observed value) are counted, and finally a p-value is obtained by dividing the total count by the number of permutations. The null hypothesis of the test is that there is no correlation between $R(X)$ and $R(Y)$. For our purposes, we chose $n = 10000$ permutations.

To produce a single value per feature per sleep recording, we averaged the feature data in the data set subject-wise. The averaging ensures that each data point carries comparable, equal weight, as the durations and hence the number of data points in the sleep recordings vary greatly. After performing subject-wise feature averaging, we computed the Spearman correlation coefficients ρ using SciPy (Virtanen et al., 2020). The results of the age-feature correlation analysis are shown in table 2.

#	**Averaged Feature**	ρ	**p-value**
1.	Movement Activity	-0.32	0.05
2.	Respiration ACF Max.	0.43	0.01
3.	Respiration Rate	-0.47	0.00
4.	Respiration Peaks Median	-0.24	0.16
5.	Respiration Peaks Std.	0.06	0.75

Table 2: Spearman correlation coefficients for the age of the subject and the subject-wise averaged feature data. Age is measured on the x-axis with months.

Moderate, low or nonexistent correlation coefficients are observed between age and all of the features, and according to the permutation tests, only the second and the third features have statistically significant correlations at the standard 5% significance level. The statistically significant observations are consistent with previous research, especially as described in the study by Fleming et al. (2011), which addressed the decrease in the respiration rate as the age increases. Positive correlation in the respiration ACF max. value feature could be linked to the development of more regular breathing patterns as infants grow older, since it is known that respiration becomes more stable and the durations and frequencies of apnoeic events during sleep reduce with age (MacLean et al., 2015).

To deepen our understanding and possibly gain further insight on the age related variations, we continued the analysis by studying the correlations on an individual sleep stage basis. In this approach, we evaluated the correlations between age and features in the different stages of sleep, which were chosen as classifier target labels, to possibly reveal development-related changes in the data in the discrete stages of sleep.

We carried out the computation of correlation coefficients and p-values in a similar fashion as in the initial analysis, but this time only data points which were identified to belong to the classes wake, N1/REM (light sleep) or N2/N3 (deep sleep) were averaged and analyzed at a time. The results are summarized in table 3.

Averaged Feature	**Wake**		**N1/REM**		**N2/N3**	
	ρ	**p-value**	ρ	**p-value**	ρ	**p-value**
Movement Activity	-0.05	0.76	0.00	0.99	0.04	0.83
Resp. ACF Max.	0.51	0.00	0.4	0.02	0.01	0.96
Respiration Rate	0.10	0.55	-0.51	0.00	-0.46	0.01
Resp. Peaks Median	-0.09	0.60	0.07	0.69	0.03	0.84
Resp. Peaks Std.	0.42	0.01	0.52	0.00	0.33	0.05

Table 3: Spearman correlation coefficients for the age of the subject and the subject-wise averaged feature data at the different stages of vigilance.

5.1.2 Regression analysis

For further understanding the relationships, we visualized the age vs. average feature data each in their own scatter plots (Figure 12) and carried out a simple regression analysis.

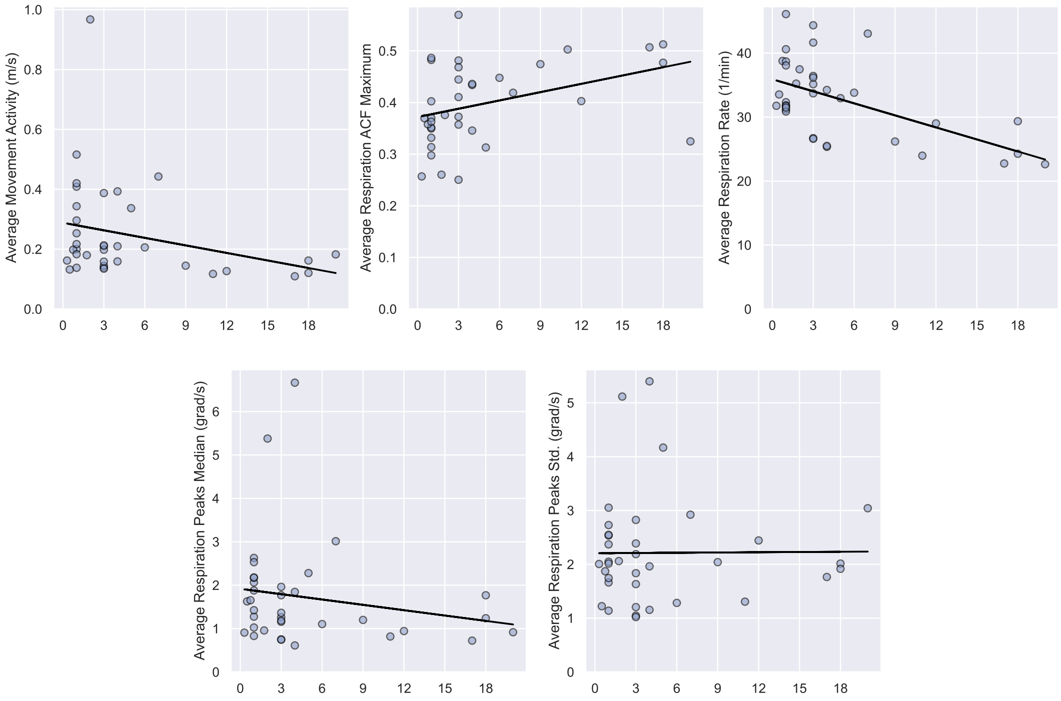


Figure 12: Subject age versus subject-wise averaged feature data.

After visually inspecting the scatter plots, we attempted to approximate the relationships as roughly linear, although the scarcity of data, high variability and outliers in the observations, and uneven distribution in ages (most of the infants were younger than six months) pose limitations to this approach. However, using linear regression, we obtained linear fits for each of the separate relationships. In theory, the regression lines (as shown in Figure 12) should allow us to predict the average feature value for an individual of a specific age. In the table below, we have gathered the regression results and goodness-of-fit (R^2) measures.

#	**Regression task**	**Slope**	**Intercept**	R^2
1.	Age vs. Movement Activity	-0.01	0.29	0.08
2.	Age vs. Resp. ACF. Max.	0.01	0.37	0.14
3.	Age vs. Respiration Rate	-0.63	35.91	0.32
4.	Age vs. Resp. Peaks Median	-0.04	1.92	0.03
5.	Age vs. Resp. Peaks Std.	0.00	2.21	0.00

Table 4: Linear fits.

5.2 Integrating age information to the model

Following the analysis of age-related correlations with the sleep-related features, we now turn our focus to evaluating whether integrating age information into our sleep classifier model can enhance its predictive performance and generalizability. Considering the impact of infant age on sleep stage indicators and characteristics, a potential challenge lies in the misclassification of sleep stages due to age-based physiological variations. This is because during training, the classifier might face difficulties associating the values of the features with the corresponding sleep stages, and according to the analysis of the previous chapter, the values appear to have dependence on the age of the infant. For example, the respiration rate indicating a certain sleep stage in a young individual might correspond to another sleep stage in an older individual. We explore two strategies in an effort to control for these age-related variations in an effort to improve the performance of the baseline classifier. The methods are described in the following subchapters.

5.2.1 Age as a separate feature

In the first approach, we directly use age as a separate sixth input feature in our classifier. Using the subject age in this manner as a separate feature could allow the model to independently to learn the relationship and adjust for age related differences in the physiological features, thereby providing an improvement to performance.

5.2.2 Age-adjusted respiration rate

In the second approach we apply age-based transformations to the features in the dataset. We used the regression fit for age versus the average respiration rate obtained in the previous section (table 4) to adjust for age-related differences between the data of the individuals. This is done by subtracting the age-estimated average values from the actual feature values in each individual sleep recording. This method only considers the respiration rate feature, because the regression fits on the other features are deemed to be insignificant due to low R^2 scores and regression slope values. This method is essentially a way of applying age-based centering to the respiration feature data, reducing the bias caused by age differences. Subtracting the age-based average values can help to highlight variations in respiration rate that are more directly related to different stages of sleep rather than age, which could improve the performance of the model. The transformation is given by:

$$y' = y - f(x).$$

Here y' are the new age corrected values for for the respiration rate, y the are original values, x is the subject age in months and the function $f(x)$ is the linear fit providing a mapping from a subject's age to the corresponding average value.

5.3 Model training

The BiGRU classifier model was developed in Python using the PyTorch framework (Paszke et al., 2019) and trained and evaluated on Lightning.ai platform equipped with a NVIDIA Tesla T4 GPU. Cross-entropy loss was used as the error metric between true labels (wake, light sleep, deep sleep) and model predictions, and the Adam optimization algorithm was adopted to minimize loss during training.

We trained each model with the following setup: learning rate of $\epsilon = 0.001$ was chosen, as lower values caused slow convergence and greater values caused unstable training and divergence in loss. We used a minibatch size of 5 full recordings, meaning that the data of 5 distinct sleep recordings were used at a time for iterative updates of the model parameters during training. As batching data to tensors requires that each sequence have the same length, the data of shorter sleep recordings were padded with extra values at the end to match the longest sleep recording in each batch (sequence length range 135 - 1199 of 30-second frames). Masking was used to ignore the values that correspond to the padding values during loss computations. For each epoch of training, the dataset was shuffled and sleep recordings were randomly selected for each minibatch. Randomly shuffling the data ensures that the model does not learn the order of the inputs, i.e., overfit to the order of the sleep recordings in the dataset.

We carried out cross-validation with the leave-one-subject-out paradigm, partitioning the dataset into 36 folds, equal to the number of individual sleep recordings. During each round of LOSOCV, the training dataset was prepared by applying z-score normalization to the features, which is a data preprocessing step that brings the mean of the data to zero and the standard deviation to one by subtracting the mean and dividing by the standard deviation. This is benecifial, as it is known that in general the convergence is faster when mean of the data is close to zero, and the features have unit variance. Normalization is also beneficial in situations where the input features have different scales, as without normalization the features with larger values could disproportionately influence the learning process, potentially leading to biased model predictions. (LeCun et al., 2012, p. 16; Jung, 2022, p. 143). The features of the test subjects were normalized by using the mean and standard deviation calculated from the training data to avoid data leakage between the training and the test data.

We used the python library TorchMetrics (Detlefsen et al., 2022), a general-purpose metrics package compatible with PyTorch, for computing the chosen measures of performance. To understand comprehensively the performance of the models, we studied the predictions of the models both on a sleep recording-level basis (computed scores and confusion matrices for each individual recording) and on individual sleep epoch basis by aggregating the model predictions from each sleep recording.

As there is inherently some randomness present in the training of deep learning models due to randomly initialized weights and stochastic optimization techniques (such as SGD or Adam), the performance of the models can vary when using different random seeds. Therefore, it is desirable to run the cross-validation and the performance measurements over multiple random seeds to minimize the confounding effect of randomness on the results. In this thesis, we used 10 different random seeds.

6 Results

6.1 Baseline Model

The results from the leave-one-out cross-validation with the original BiGRU model are presented in this section. Evolution of the classifier performance during training is visualized in Figure 13, which shows the averaged learning curves over the whole cross-validation procedure. The test and training loss decrease with the same pace up around 20 epochs, after which the training loss continues to decrease faster than the test loss. The average test loss stagnates after 100 epochs, while training loss continues to decrease. The model begins to overfit, and training is halted.

The classifier achieved a median sleep recording-level accuracy of 77% and a median sleep recording-level MCC score of 0.64. The standard deviation for accuracy was 9% and 0.13 for MCC. Aggregated epoch-to-epoch accuracy was 76% and MCC 0.62. Class-wise record-level median accuracies: 90% (Deep sleep), 70% (Light sleep), 80% (Wake). Spearman correlation between infant age and MCC score was -0.31.

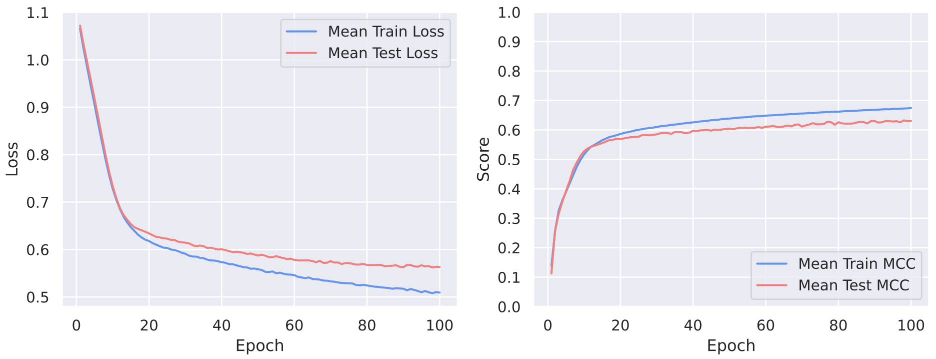
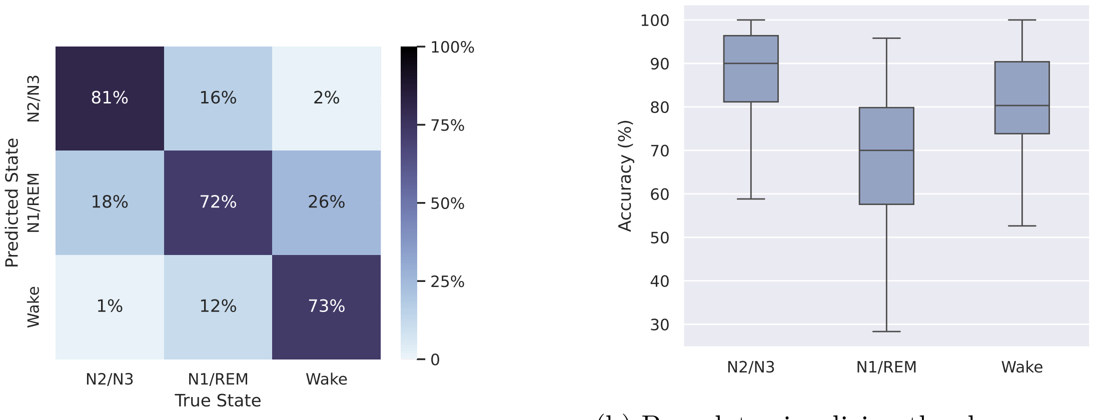


Figure 13: Learning curves depicting the performance of the classifier as a function of training epochs. Left: Average Cross-Entropy Loss. Right: Average MCC.



(a) Confusion matrix of the aggregated model predictions vs. true labels.

(b) Box plots visualizing the sleep recording level accuracies for different sleep stages.

Figure 15 shows the hypnogram produced by the BiGRU classifier superimposed on the physician-generated hypnogram from a PSG sleep study, and the classifier confidence, representing the likelihood of the most probable class. The classifier confidence is lowest during periods with frequent sleep stage transitions and most confident during longer periods of a single sleep state.

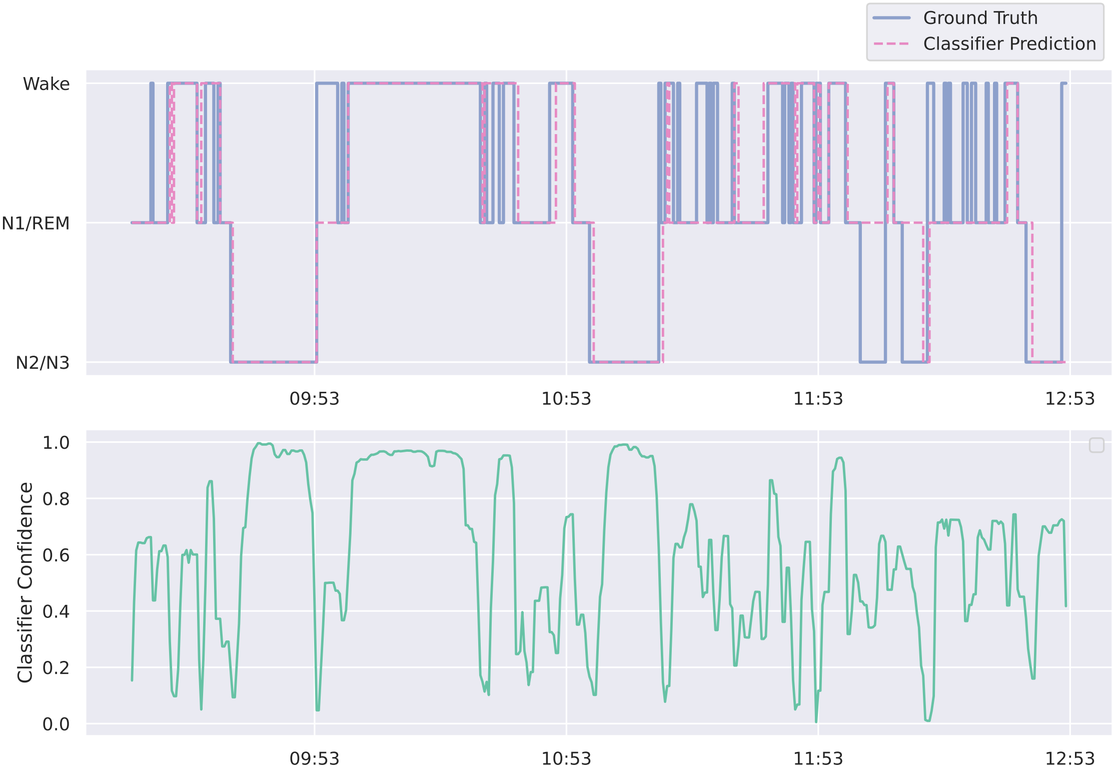
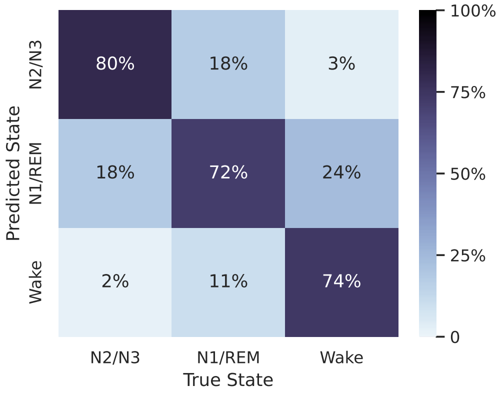


Figure 15: Time series plots illustrating the PSG derived hypnogram (ground truth), the artificially generated hypnogram (classifier prediction) and classifier confidence (linearly scaled to interval [0, 1]).

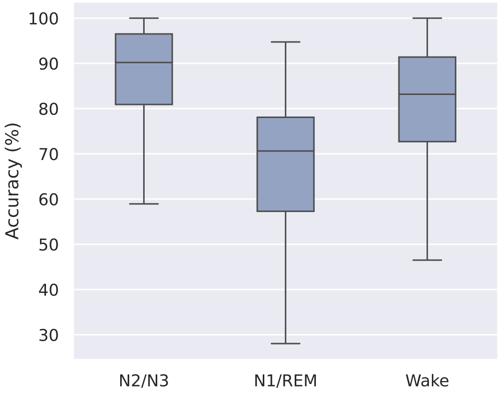
6.2 Integrating age information to the model

6.2.1 Age as separate feature

The results from the model with age as a separate feature are presented in this section. The classifier achieved a median accuracy of 78% and a median MCC score of 0.64. The standard deviation for accuracy was 9% and 0.13 for MCC. Aggregated epoch-to-epoch accuracy was 76% and MCC 0.62. Class-wise record-level median accuracies: 90% (Deep sleep), 71% (Light sleep), 83% (Wake). Spearman correlation between infant age and MCC score was -0.32.



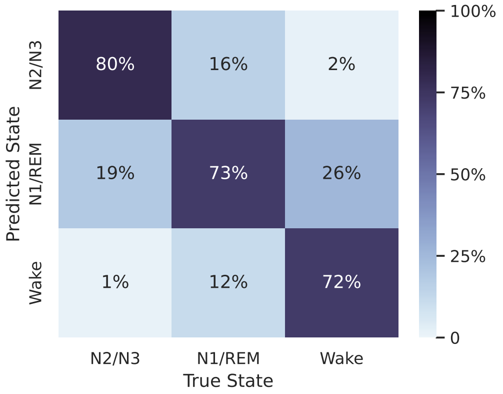
(a) Confusion matrix of the aggregated model predictions vs. true labels.



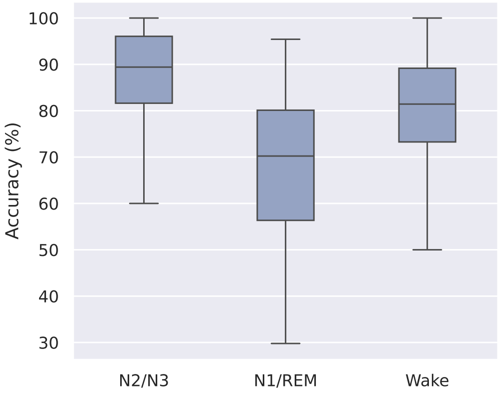
(b) Box plots visualizing the sleep recording level accuracies for different sleep stages.

6.2.2 Age-adjusted respiration rate

The classifier with an age-adjusted respiration rate achieved a median accuracy of 77% and a median MCC score of 0.64. The standard deviation for accuracy was 9% and 0.13 for MCC. Class-wise record-level median accuracies: 89% (Deep sleep), 70% (Light sleep), 81% (Wake). Correlation between infant age and MCC score was -0.34.



(a) Confusion matrix of the aggregated model predictions vs. true labels.



(b) Box plots visualizing the sleep recording level accuracies for different sleep stages.

7 Discussion

The model introduced in this thesis uses wearable sensor-based respiratory and movement-related features for classification of sleep stages. On both the level of sleep recordings and individual sleep epochs, the model achieved good performance scores that are comparable to the results of previous research, discussed in chapter 2, although research is lacking in the field of automatic sleep scoring of infant sleep using wearable technology. The baseline model without age-related adjustments achieved a median accuracy score of 77% and an MCC score of 0.64 in classifying the merged sleep stages of wake, light sleep (N1/REM) and deep sleep (N2/N3). At the aggregated epoch-to-epoch level, the baseline model achieved an accuracy of 76% and a MCC of 0.62, both values being close to the record-level scores.

Although the scores obtained can be considered good, the LOSOCV procedure yielded somewhat high standard deviations for the scores in all experiments (9% for accuracy and 0.13 for MCC), which is usually a common sign of overfitting (Goodfellow et al., 2016, p. 130). However, the high variability can likely be attributed to individual differences between sleep recordings, rather than being an indicator of an ill-posed model. This is because when we tried using common steps to mitigate overfitting in deep learning (such as reducing model complexity, L^2 regularization, and early stopping), sleep recordings with low performance tended to stay low while sleep recordings with high accuracy tended to stay high. This high variability within performance scores probably originates from the fact that the recordings in the dataset are very heterogeneous in many different aspects. For example, differences in infant health (the study recruits suffered from various sleep related medical disorders), sleep recording lengths, high imbalance of sleep stages and quick and frequent transitions between sleep stages in the sleep recordings introduce challenges for the classifier. We noted that the classifier had particularly low confidence and poor performance in the transitions between sleep stages. Therefore, if a particular recording had frequent transitions, this was reflected in the performance scores. Furthermore, the light sleep class (N1/REM) had on average the poorest recognizability (this is seen from the box plots in the previous section, where the box plot of light sleep class has particularly wide whiskers.). Also, the light sleep class has considerable overlap and egde cases with the adjacent deep sleep and wake classes, which can be seen in the the t-SNE visualization of the sleep stages (Figure 9). According to confusion matrices, the classifier mistakenly interprets the wake state as the light sleep state in 25% of the cases, a considerably high portion. The high portion of light sleep/wake in a sleep recording could thus be reflected as a lower classifier performance.

The main objective of this thesis was to study whether infant age had an impact to the input features and model performance, and consequently if incorporating age-related information into the sleep classifier model could improve the classification performance. Indeed, we found that there was a moderate relationship between age and average respiration rate ($\rho \approx -0.5$), with other features displaying lower correlations. We found that neither of the strategies researched in this thesis turned out to be helpful in mitigating the discrepancy caused by variability of infant age in the classifier performance. The baseline model showed a negative correlation between infant age and MCC score of -0.31, and interestingly enough this negative correlation was slightly increased when we incorporated the age adjustments to the model. When we used age as a separate feature, this correlation was -0.32 and -0.34 when respiratory rates were adjusted for age. The median performance scores at sleep recording level had little to no difference between the modeling approaches, and no significant differences were observable in the aggregated confusion matrices. The class-wise recording-level accuracies of the sleep stages showed the most difference between the models; The median accuracy for the wake class increased from 80% to 83% when age was used as a separate feature. Based on these results, it is evident that the modeling approaches we introduced in this thesis did not contribute significantly to the performance of the classifier. As the BiGRU model we employed takes into consideration both the future and history of every time step in a given sleep recording, it might be possible that the model learns itself to correct for the age-related differences.

The main limitations in our research was the scarcity of data (N = 36), short sleep recording spans (median length only 4 hours) and the fact that the recruits of the sleep study cohort were suffering from various sleep related medical conditions. Furthermore, older infants were highly underrepresented in the dataset, as half of the infants were younger or at most as old as 3 months. Only 8 of the infants were older than 6 months. Due to the small number of sleep recordings in the dataset, we did not follow the conventionally used train, validation and test splitting procedure to evaluate the classifier performance, but rather only used the LOSOCV paradigm. For fully optimizing the training and model hyperparameters and thus potentially obtaining better performance results, the cross-validation should be carried out on a separate validation set for hyperparameter selection, and model evaluation should be done on a completely separate test set.

This thesis paves the way for additional research. Further methods in infant sleep stage classification could explore alternative, more advanced deep learning models which have proven to be successful in other time series analysis tasks, such as the encoder-decoder model and the transformer, which has been the main driving force behind the recent AI boom. Alternative approaches to age-adjusted modeling could include an ensemble model, where multiple models are trained and tested against different binned age groups, and then lastly combined. One model in such ensemble could be trained, for example, on the data from infants aged 0 up to 3 months, the next on ages from 3 to 6 months, etc. An ensemble approach could help to improve the performance, as the classifier weights do not have to generalize over data with large variance in age. Another technique that could be explored is a custom loss function, which weights the loss based on the age of the infant, giving larger penalty scores for older infants. Multitask learning could be leveraged in an approach where the model is simultaneously taught to predict the infant age based on the input features, while doing the sleep staging. The models could then leverage shared information between them to improve performance in both tasks. To enhance the reliability of the results' interpretation, employing statistical significance tests could be used to determine whether the variations observed in the modeling techniques stem from randomness or from genuine disparities in classifier performance. Appropriate test for such case could be the Wilcoxon signed-rank test.

8 Conclusion

In this thesis, we sought to improve the performance of a Gated Recurrent Unit (GRU) based sleep stage classifier in infant populations using wearable technology. Building upon the work of Ranta et al. (2021a), our primary focus revolved around the NApping PAnts (NAPPA) system, a wearable device designed for infants utilizing a gyroscope and an accelerometer for monitoring movement activity and respiratory related movements during sleep. This system was developed for sleep monitoring purposes in the outside-of-hospital setting, allowing easy, cost-effective, and precise monitoring of infant sleep health in long-term studies. Combined with the GRU-based machine learning classifier, the NAPPA system yielded good overall results in automatic sleep stage classification.

However, as infant sleep is highly dynamic and sleep-related physiological indicators, such as respiration rate and stability change as infants grow, research was needed to see if accounting for these changes in the classifier model could further improve the performance of the classifier.

Using a dataset of 36 sleep recordings from distinct infants aged 2 weeks to 20 months, we first investigated the relationship between infant age and the various respiratory related features that were used as inputs to the BiGRU classifier. This analysis was carried out by computing Spearman correlation coefficients (ρ) between infant ages and the averaged feature values from the NAPPA sensor data in the sleep recordings. We further investigated the relationships by computing the correlations at individual stages of vigilance (Wake, Light Sleep, and Deep Sleep) to gain more understanding as to how age could affect the various respiratory phenomena in these different states. We found through this analysis that the average respiration rate has a negative correlation ($\rho \approx -0.5$) with age and decreases on average by 8 breaths per minute per year, a finding in alignment with previous research conducted by Fleming et al. (2011).

We provided visualizations of the relationships and used linear regression to obtain linear fits mapping infant age to average feature values, which brings us to the main concern of this thesis. That is, through our methods of data preprocessing and feature engineering, we sought to make the portable NAPPA system more adaptive to the varying physiological factors, such as the decreasing respiratory rate, seen across different age groups in this population. The main methods assesssed in making of this study included the incorporation of age-adjusted feature transformations by using the linear fit mapping infant age to average respiration rate value. This mapping was used to subtract age-estimated average respiration rate values from the sleep recording data in an attempt to center the data by age, an approach in which the purpose was to enable the classifier to focus to changes in respiration between sleep stages, rather than the changes caused by different ages of the infants in the dataset. In our second approach, we modified the BiGRU model architecture so that we added the infant ages as separate input features alongside the respiratory features, a goal in which we sought to train the model independently correct for the age-causing discrepansies in the feature data. Despite our concerted efforts, neither approach yielded significant improvements in the classifier's performance. Our findings underscore the complexity of infant sleep classification and the challenges associated with mitigating the impact of age-related physiological variations on classifier performance. Moreover, our study highlights the need for further research to explore alternative modeling approaches and address the inherent complexities of infant sleep data. Future research endeavors could explore alternative deep learning architectures, such as encoder-decoder models and transformers, to tackle the intricacies of infant sleep classification. Moreover, the adoption of ensemble modeling techniques, tailored to specific age groups, may offer a promising avenue for improving classifier performance. In conclusion, while our study did not yield the desired enhancements to the GRU-based sleep stage classifier, it sheds light on the multifaceted nature of infant sleep classification and underscores the importance of continued research in this domain. By addressing the methodological challenges and refining modeling approaches, we can strive towards more robust and accurate solutions for monitoring infant sleep health using wearable technology.

References

J. Liu, X. Ji, S. Pitt, G. Wang, E. Rovit, T. Lipman, and F. Jiang. Childhood sleep: physical, cognitive, and behavioral consequences and implications. *World Journal of Pediatrics: WJP*, 20(2):122–132, 2024.

E. Mervaala, E. Haaksiluoto, S.L. Himanen, S. Jääskeläinen, M. Kallio, and S. Vanhatalo. *Kliininen neurofysiologia*, volume 1. Duodecim, Helsinki, 2019.

O. Bruni and L. Novelli. Sleep disorders in children. *BMJ clinical evidence*, 2010.

E. K. Tham, N. Schneider, and B. F. Broekman. Infant sleep and its relation with cognition and growth: a narrative review. *Nature and science of sleep*, 9:135–149, 2017.

Y. J. Lee, J. Y. Lee, J. H. Cho, and J. H. Choi. Interrater reliability of sleep stage scoring: a meta-analysis. *Journal of Clinical Sleep Medicine: JCSM*, 18(1): 193–202, 2022.

A. Sadeh. Sleep assessment methods. *Monographs of the Society for Research in Child Development*, 80(1):33–48, 2015.

J. Ranta, E. Ilén, K. Palmu, J. Salama, O. Roienko, and S. Vanhatalo. An openly available wearable, a diaper cover, monitors infant's respiration and position during rest and sleep. *Acta Paediatrica*, 110(10):2766–2771, 2021a.

K. Cho, B. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 1724–1734, 2014.

S. de Sena, M. Häggman, J. Ranta, O. Roienko, E. Ilén, N. Acosta, J. Salama, T. Kirjavainen, N. Stevenson, M. Airaksinen, and S. Vanhatalo. Napping pants (nappa): An open wearable solution for monitoring infant's sleeping rhythms, respiration and posture. *Heliyon*, 10(13):e33295, 2024.

A. K. Patel, V. Reddy, K. R. Shumway, and J. F. Araujo. Physiology, sleep stages. *Website*, 2022. . Cited 15.09.2023. Available: https://www.ncbi.nlm.nih.gov/books/NBK526132/.

J. E. MacLean, D. A. Fitzgerald, and K. A. Waters. Developmental changes in sleep and breathing across infancy and childhood. *Paediatric Respiratory Reviews*, 16(4):276–284, 2015.

A. R. Tarullo, P. D. Balsam, and W. P. Fifer. Sleep and infant learning. *Infant and child development*, 20(1):35–46, 2011.

S. Fleming, M. Thompson, R. Stevens, C. Heneghan, A. Plüddemann, I. Maconochie, L. Tarassenko, and D. Mant. Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies. *Lancet*, 9770(377):1011–1018, 2011.

A. DeMasi, M. N. Horger, A. Scher, and S. E. Berger. Infant motor development predicts the dynamics of movement during sleep. *Infancy*, 28(2):367 – 387, 2023.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(9): 273–297, 1995.

I. Aboalayon, M. Faezipour, S. Almuhammadi, and S. Moslehpour. Sleep stage classification using eeg signal analysis: A comprehensive survey and new investigation. *Entropy*, 18(9), 2016.

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

R. N. Sekkal, F. Bereksi-Reguig, D. Ruiz-Fernandez, N. Dib, and S. Sekkal. Automatic sleep stage classification: From classical machine learning methods to deep learning. *Biomedical Signal Processing and Control*, 77:103751, 2022.

N. J. Douglas, D. P. White, , C. K. Pickett, J. V. Weil, and C. W. Zwillich. Respiration during sleep in normal man. *Thorax*, 37(11):840–4, 1982.

G. Haddad, H. Jeng, T. Lai, and R. Mellins. Determination of sleep state in infants using respiratory variability. *Pediatric Research*, 21(8):556–562, 1987.

R. Harper, V. Schechtman, and K. Kluge. Machine classification of infant sleep state using cardiorespiratory measures. *Electroencephalography and Clinical Neurophysiology*, 67(4):379–387, 1987.

G. Z. Liu, Y. W. Guo, Q. S. Zhu, B. Y. Huang, and L. Wang. Estimation of respiration rate from three-dimensional acceleration data based on body sensor network. *Telemedicine journal and e-health : the official journal of the American Telemedicine Association*, 17(9):705–711, 2011.

F. Ryser, S. Hanassab, O. Lambercy, E. Werth, and R. Gassert. Respiratory analysis during sleep using a chest-worn accelerometer: A machine learning approach. *Biomedical Signal Processing and Control*, 78, 2022.

M. O. Mendez, M. Migliorini, J. M. Kortelainen, D. Nistico, E. Arce-Santana, S. Cerutti, and A. M. Bianchi. Evaluation of the sleep quality based on bed sensor signals: Time-variant analysis. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society.*, page 3994–3997, 2010.

J. M. Kortelainen, M. O. Mendez, A. M. Bianchi, M. Matteucci, and S. Cerutti. Sleep staging based on signals acquired through bed sensor. *IEEE transactions on information technology in biomedicine*, 14(3):776–785, 2010.

A. Tal, Z. Shinar, D. Shaki, S. Codish, and A. Goldbart. Validation of contact-free sleep monitoring device with comparison to polysomnography. *Journal of clinical sleep medicine*, 13(3):517–522, 2017.

M. Gaiduk, T. Penzel, J. A. Ortega, and R. Seepold. Automatic sleep stages classification using respiratory, heart rate and movement signals. *Physiological measurement*, 39(12):124008, 2018.

Y. Zhang, Z. Yang, K. Lan, X. Liu, Zhengbo Zhang, P. Li, D. Cao, J. Zheng, and J. Pan. Sleep stage classification using bidirectional lstm in wearable multi-sensor systems. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 443–448, 2019.

J. Ranta, M. Airaksinen, T. Kirjavainen, S. Vanhatalo, and N. J. Stevenson. An open source classifier for bed mattress signal in infant sleep monitoring. *Frontiers in Neuroscience*, 14, 2021b.

A. Tataraidze, L. Anishchenko, L. Korostovtseva, B.J. Kooij, M. Bochkarev, and Y. Sviryaev. Sleep stage classification based on respiratory signal. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 358–361, 2015.

J. Yang, J. M. Keller, M. Popescu, and M. Skubic. Sleep stage recognition using respiration signal. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society.*, page 2843–2846, 2016.

M. Baumert, S. Hartmann, and H. Phan. Automatic sleep staging for the young and the old – evaluating age bias in deep learning. *Sleep Medicine*, 107:18–25, 2023.

V. L. Schechtman and R. M. Harper. The maturation of correlations between cardiac and respiratory measures across sleep states in normal infants. *Sleep*, 15 (1):41–47, 1992.

G. Litscher, G. Pfurtscheller, F. Bes, and E. Poiseau. Respiration and heart rate variation in normal infants during quiet sleep in the first year of life. *Klinische Padiatrie*, 205(3):170–175, 1993.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. URL http://www.deeplearningbook.org.

Maximilian Schreiner. Gpt-4 architecture, datasets, costs and more leaked. *Website*, 2023. . Cited 12.5.2024. Available: https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/.

M. Leshno, V. Ya. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

A. Jung. *Machine Learning: The Basics*. Springer, Singapore, 2022.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. URL https://arxiv.org/abs/1412.6980.

S. Linnainmaa. Algoritmin kumulatiivinen pyöristysvirhe yksittäisten pyöristysvirheiden taylor-kehitelmänä. Master's thesis, University of Helsinki, Helsinki, 1970.

D. Rumelhart, G. Hinton, and R. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

S. Yang, X. Yu, and Y. Zhou. Lstm and gru neural network performance comparison study: Taking yelp review dataset as an example. In *2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)*, pages 98–101, 2020.

R. Cahuantzi, X. Chen, and S. Güttel. A comparison of lstm and gru networks for learning symbolic sequences. In *Intelligent Computing*, pages 771–785, Cham, 2023. Springer Nature Switzerland.

B.W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451, 1975.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Metrics and scoring: quantifying the quality of predictions. *Website*, 2021. . Cited 19.04.2024. Available: https://scikit-learn.org/stable/modules/model_evaluation.html#matthews-corrcoef.

W. Mckinney. pandas: a foundational python library for data analysis and statistics. *Python High Performance Science Computer*, 01 2011.

J. N. Acosta-Leinonen. Monitoring newborn and infant sleep respiration and heart rate with a wearable sensor. Master's thesis, University of Helsinki, Helsinki, 2019.

G. E. Hinton and S. Roweis. Stochastic neighbor embedding. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002.

D. L. Mohr, W. J. Wilson, and R. J. Freund. *Statistical Methods*. Academic Press, fourth edition edition, 2022. ISBN 978-0-12-823043-5.

P. Virtanen, R. Gommers, and T. E. et al. Oliphant. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272, 2020.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E.Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703, 2019.

Y. A. LeCun, B. Léon, G. B. Orr, and K. R. Müller. *Efficient backprop*, pages 9–48. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

N. Detlefsen, J. Borovec, J. Schock, A. Jha, T. Koker, L. Di Liello, D. Štancl, Q. Changsheng, M. Grechkin, and W. Falcon. Torchmetrics -measuring reproducibility in pytorch. *The Journal of Open Source Software*, 7, 2022.