Tampereen yliopisto

Johanna Kujala

PERFORMANCE EVALUATION OF DEEP LEARNING TOOLS FOR SLEEP STRUCTURE ASSESSMENT IN CLINICAL SETTINGS

Master of Science (Tech.) Thesis
Faculty of Medicine and Health Technology
Supervisors and Examiners: Mark van Gils
Antti Kulkas
Sami Nikkonen
Antti Kallonen
April 2025

ABSTRACT

Johanna Kujala: “Performance evaluation of deep learning tools for sleep structure assessment in clinical settings”
Master’s thesis
Tampere University
Master's Programme in Biomedical Sciences and Engineering
Supervisors: Chief Physicist, Adjunct professor Antti Kulkas, Adjunct professor Sami Nikkonen, and Professor Mark van Gils
April 2025

Accurate sleep assessment is crucial for diagnosing sleep disorders and monitoring patient health. Clinically, sleep is commonly assessed using overnight polysomnography (PSG) recordings, which provide several hours of data from multiple channels. The data analysis, conducted by-hand, is a labor-intensive, time-consuming and expensive task requiring expert knowledge, limiting the availability of the measurements and scalability of the analysis processes. This has led to active research into the possibilities to automate the analysis process. Deep learning has been identified as one possible avenue for automating the analysis of sleep.

Conventionally, sleep and sleep structure are assessed using electroencephalography (EEG), electro-oculography and submental electromyography (EMG) signals. Based on these signals, sleep is clinically classified in five stages: wake (W), REM-sleep (R) and N1, N2 and N3 sleep. The photoplethysmogram (PPG) signal has been explored as an alternative signal for assessing sleep using deep learning -based tools. Analysis based on the PPG signal could be especially beneficial in situations where conventional channels for sleep assessment are not available.

In this study, two trained deep learning models were employed for sleep staging in an independent clinical PSG dataset. The first PSG model utilized one EEG channel (F4-M2), one EOG channel (E1-M2) and PPG waveform, nasal pressure, blood oxygen saturation, and respiratory inductance plethysmography (RIP) sum signals in the analysis, whereas the second polygraphy (PG) model used PPG waveform, nasal pressure, saturation, and RIP sum signals. To assess the model performance in the clinical dataset the analysis results of the deep learning models were compared to the analysis conducted by a medical doctor specialized in sleep medicine.

The accuracy in the five-stage sleep classification was 71.7 % for the PSG model analysis and 60.9 % for the PG model analysis. The respective Cohen’s kappa values were 0.58, and 0.45. Only minor differences were observed in the model performances between male and female subjects. For both models, a decrease in overall performance was observed with increasing obstructive sleep apnea (OSA) severity. In addition to the five-stage sleep assessment, the ability to distinguish between wakefulness and sleep was evaluated. Generally, both models were better able to correctly identify sleep compared to wakefulness.

Although deep learning -based methods have shown promising results in sleep classification, this work demonstrates that trained deep learning models may not perform consistently across independent datasets. The results also suggest that OSA severity might impact model performance, possibly due to increased sleep fragmentation associated with more severe OSA. Further research is warranted to ensure the applicability of these methods across different datasets as well as to understand the factors affecting performance variations for individuals across these datasets.

Keywords: sleep, sleep architecture, deep learning

TIIVISTELMÄ

Johanna Kujala: “Syväoppimismenetelmien suorituskyvyn arviointi unen rakenteen analysoinnissa kliinisessä ympäristössä”
Diplomityö
Tampereen yliopisto
Bioteknologian ja biolääketieteen tekniikan tutkinto-ohjelma
Työn ohjaajat: Ylifyysikko, Dosentti Antti Kulkas, Dosentti Sami Nikkonen ja Professori Mark van Gils
Huhtikuu 2025

Unen ja unen rakenteen arviointi on olennainen osa unihäiriöiden diagnostiikkaa ja potilaiden terveyden seurantaa. Kliinisesti unta mitataan läpi yön tyypillisesti unipolygrafialla (PSG). Nämä mittaukset tuottavat useita tunteja dataa useasta eri mittauskanavasta. Mittaustulokset käydään käsin läpi kliinisen ammattilaisen toimesta, mikä tekee prosessista työlään, aikaa vievän ja kalliin. Tämä rajoittaa myös mittausten saatavuutta, eikä analyysiprosessi ole skaalautuva. Manuaaliseen analyysiin liittyvät rajoitteet ovat innoittaneet aktiivista tutkimustyötä prosessin automatisoimiseksi. Syväoppimispohjaiset analyysimenetelmät ovat yksi mahdollinen lähestymistapa unen analyysin automatisoimiseksi.

Perinteisesti unen rakenteen arvioinnissa hyödynnetään elektroenkefalografia (EEG), elektrookulografia (EOG) ja leuka-elektromyografia (EMG) kanavista saatavaa informaatiota. Näiden signaalien sisältämän informaation perusteella uni luokitellaan kliinisesti viiteen eri luokkaan: valveeseen (W), REM-uneen (R) sekä N1-, N2- ja N3-uneen. Fotopletysmografiasignaalia (PPG) on tutkittu mahdollisena vaihtoehtoisena kanavana syväoppimistyökaluihin perustuvassa unen arvioinnissa. PPG-signaaliin pohjautuva analyysi voitaisiin hyödyntää erityisesti tilanteissa, joissa perinteisiä unen arvioinnissa käytettäviä kanavia ei ole tarjolla.

Tässä työssä kahta entuudestaan koulutettua syväoppimispohjaista koneoppimismallia käytettiin unen rakenteen analysoimiseksi itsenäisessä kliinisessä unipolygrafia-aineistossa. Ensimmäinen malli (PSG-malli) käyttää analyysissa yhtä EEG-kanavaa (F4-M2), yhtä EOG-kanavaa (E1-M2) sekä PPG-, nenäpaine-, happisaturaatiotaso- ja RIP-signaaleja. Toinen polygrafia-malli (PG-malli) käyttää analyysissa PPG-, nenäpaine-, happisaturaatiotaso- ja RIP-signaaleja. Mallien analyysituloksia verrattiin unilääketieteeseen erikoistuneen lääkärin tekemään analyysiin.

Koko aineistossa viiden univaiheen luokittelussa PSG-mallin tarkkuus oli 71,7 % ja PG-mallin tarkkuus 60,9 %. Vastaavat Cohenin kappa -arvot olivat 0,58 ja 0,45. Mallin suorituskyvyssä miesten ja naisten välillä havaittiin vain pieni eroja. Molemmilla malleilla suorituskyky laski asteittain uniapnean vaikeusasteen kasvaessa. Viiden univaiheen luokittelukyvyn lisäksi arvioitiin tekoälymallien kykyä erotella uni ja valve toisistaan. Mallit olivat keskimäärin parempia tunnistamaan oikein unta valveeseen verrattuna lähes kaikissa tutkituissa potilasryhmissä.

Vaikka syväoppimispohjaisilla menetelmillä on saatu lupaavia tuloksia univaiheiden luokittelussa, tämä työ osoittaa, että tähän tarkoitukseen koulutettujen syväoppimismallien suorituskyvyssä saattaa olla vaihtelua eri kliinisissä aineistoissa. Tämän tutkimuksen tulokset antavat lisäksi viitteitä siitä, että mallin suorituskyky saattaa laskea uniapnean vaikeusasteen kasvaessa. Tämä voi johtua siitä, että vaikeammassa uniapneassa unen rakenne on tyypillisesti katkonaisempaa. Lisätutkimuksia tarvitaan, jotta näiden menetelmien sovellettavuus eri aineistoihin voidaan kehittää. Lisäksi lisää ymmärrystä tarvitaan mallin suorituskykyyn vaikuttavista tekijöistä yksittäisillä potilailla.

Avainsanat: uni, unen rakenne, syväoppiminen

USE OF ARTIFICIAL INTELLIGENCE IN THIS THESIS

Artificial intelligence (AI) has been used in generating this work:

☐ No
☒ Yes

I hereby declare that the AI-based applications used in generating this work are as follows:

Application	Version
ChatGPT	GPT-40
PaperPal	2.16.3

Purpose of the use of AI:

ChatGPT was used to help form some of the python scripts for preprocessing the data, plotting the confusion matrices and calculating the performance metrics. Additionally, ChatGPT was utilized to correct the grammar or improve the flow of some individual sentences or parts of sentences. A grammar check for the entire work was performed using PaperPal.

Parts of this work, where AI was used:

AI was used while writing the Python scripts for data processing and visualization. During the writing process, grammatical errors and flow of sentences were corrected based on suggestions of the AI tools when seen applicable. Such corrections were applied in all sections of the written work.

I hereby acknowledge that as the author of this work, I am fully responsible for the contents presented in this thesis. This includes the parts that were generated by an AI, in part or in their entirety. I therefore also acknowledge my responsibility in the case where the use of AI has resulted in ethical guidelines being breached.

PREFACE

Few things in life are achieved alone, and this thesis was no exception. Gratitude should be placed where it belongs, and a great portion of it definitely belongs to my supervisors Antti Kulkas, Sami Nikkonen and Mark van Gils. First, an immense thank you to Antti for the opportunity to work in such an inspiring environment, and for the close guidance and patience throughout this thesis process. The green light after a knock on your door has become very familiar during the past months, and I don't recall a question or a problem that was not somehow at least helped forward. Thank you, Sami, for all the help with the tools and data, and the thorough and thoughtful comments on the work and other matters. Although Kuopio was a safe distance away regarding knocks on the door, the distance neither spared you from my questions nor diminished the value of the insights you provided along the way. And thank you, Mark, for the timely responses to questions, the valuable help with troubleshooting, and feedback on the work.

This master's thesis was conducted for the Wellbeing Services County of South-Ostrobothnia and carried out at the Department of Clinical Neurophysiology at Seinäjoki Central Hospital. The work was performed in close collaboration with the Sleep Research and Analytics Group (STAR). I would like to extend my gratitude to the staff of the department of Clinical Neurophysiology in Seinäjoki Central Hospital for making the working environment so welcoming and professional, and to the staff of the STAR group for the welcoming and helpful atmosphere. Special thanks to Miika Vainikka from STAR group for the help with implementing the tools and for the willingness to bounce back and forth ideas during the process.

Thank you also to my family and friends for the practical help, the listening ears and the moments of connection – for more than can be written here. Especially thanks to my partner Aarni for the small portions of caring, silliness and fun you incorporate into every day. And finally, thanks to my horse Memo who, if need be, will get stuck in a fence to remind me that the world, in fact, revolves around them (not work or anything else). Although it must be noted that these thanks are probably appreciated only in an edible form. I am uncertain how good this thesis would taste, so maybe a carrot will do.

Seinäjoki, 23.4.2025

Johanna Kujala

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

AASM	American Academy of Sleep Medicine
AC	Alternating Current
AHI	Apnea-Hypopnea Index
AI	Artificial Intelligence
CNN	Convolutional Neural Network
DC	Direct Current
EEG	Electroencephalogram
EMG	Electromyogram
EOG	Electrooculogram
FN	False Negatives
FP	False Positives
f_s	Sampling rate (Sampling frequency)
GAP	Global Average Pooling
HF-AC	High Frequency Alternating Current
IQR	Inter-Quartile Range
LF-AC	Low Frequency Alternating Current
OSA	Obstructive Sleep Apnea
PG	Polygraphy
PPG	Photoplethysmography
PSG	Polysomnography
ReLU	Rectified Linear Unit
REI	Respiratory Event Index
REM	Rapid Eye Movement
RIP	Respiratory Inductance Plethysmography
RNN	Recurring Neural Network
TN	True Negatives
TP	True Positives

.

1. INTRODUCTION

Sleep is a cornerstone of physical, mental, and emotional well-being, and plays a critical role in maintaining overall health. Despite the lack of conscious awareness during sleep, the brain remains highly active in processing information, consolidating memories, removing metabolic products and toxins from the brain, and promoting physical and physiological recovery [1]. Insufficient or poor-quality sleep is associated with significant adverse effects, including fatigue, cognitive impairment, mood dysregulation, and hormonal and immune system dysfunction [1].

Obstructive sleep apnea (OSA) is a common sleep-related breathing disorder known to negatively affect sleep quality. OSA is estimated to affect up to 9-38 % of the general population [2]. Individuals suffering from OSA experience several complete or partial cessations of breathing throughout the night, often resulting in fragmentation of sleep as an effort to restore normal breathing [3], [4]. OSA diagnosis and severity assessment is mainly based on the average number of these cessations per hour of sleep [3], [4].

Polysomnography (PSG) measurements provide information on sleep structure and sleep-related breathing disturbances, and it is considered the gold standard for evaluating sleep and identifying sleep disorders [5] as well as for diagnosing OSA [4]. The PSG data collected during the entire night is reviewed and analyzed by hand by qualified hospital personnel, a process that may take several hours [5]. This resource-intensive and time-consuming procedure contributes to the limited availability of PSG recordings. To address these issues, the possibilities of utilizing deep learning to automate the classification of sleep stages [5], [6], [7], [8], [9], [10] and the detection of respiratory events [11], [12] has become an active field in the research community. A diagnostic alternative for OSA is ambulatory cardiorespiratory polygraphy (PG) measurement [1]. Like PSG measurements, PG measurements also include cardiorespiratory monitoring, allowing for the detection of breathing cessations. However, PG does not include the conventional channels used for sleep assessment. Although PG measures are less intrusive, easier to perform, and more cost-effective, the lack of information on sleep structure can lead to a less accurate assessment of the nature and severity of diseases such as OSA [13]. Recent efforts to determine sleep stages from photoplethysmography (PPG) measurements have attempted to address this limitation [11], [14], [15], [16], opening new possibilities for sleep assessment without relying on EEG monitoring.

Despite some promising results for automated sleep staging, automatic scoring is still not in widespread clinical use [1], [5], indicating a need for further research to enhance the accuracy, reliability, and clinical acceptance of these methods. This thesis contributes to the research by using two different deep learning models to analyze sleep structure. The first model was developed to analyze selected signals from PSG measurements, whereas the second model was designed to operate on the more limited signal set available in both PSG and PG recordings. The aim of this work was to evaluate the performance of these previously developed and trained deep learning models in an independent clinical PSG dataset recorded from patients with suspected OSA or other sleep-related pathophysiologies. Furthermore, we aimed to evaluate how subject sex or OSA severity affect the models' performance in this dataset. Finally, we aimed to evaluate the usability and suitability of deep learning methods in a clinical context. The existing literature has shown that models with similar architecture have reached accuracies between 69–85 % using various input channels [7], [9], [11] and the original test data analysis on the performance of our models is in line with these results with accuracies of 84 % and 74 %. However, acknowledging possible limitations in generalization of the models, we hypothesized that the models' performances would decrease when transferred to this new independent dataset. Additionally, we hypothesized that the model performance will decrease with increase in OSA severity.

In the following sections, we begin by delving into the theoretical background, including sections discussing sleep and sleep architecture, OSA and its effects on sleep, and the methods used to measure sleep and nighttime breathing. The focus is then shifted to deep learning, with an emphasis on the types of neural networks relevant to this study. The chapter is concluded with a review of previous research on the use of deep learning to analyze sleep architecture. The materials and methods section describes the data collection and processing, computing environments and deep learning models used in this study, and processing of the automated analysis results. The findings are then presented in the results section and further contextualized in the discussion section, where their broader implications and directions for future research are also explored. Finally, a brief summary is presented in the conclusions section.

2. THEORETICAL BACKGROUND

This chapter provides the theoretical background necessary to understand the key concepts underlying this study. First, clinical foundations are established by describing how sleep is categorized into distinct stages that can be identified based on physiological signals. This is followed by an introduction to OSA, a prevalent sleep-related breathing disorder that disrupts normal sleep patterns, as this study includes an evaluation of model performance across patient subgroups defined by OSA severity. The discussion then expands to cover measurement techniques for both sleep and respiratory activity, providing essential context for understanding the data used in this study. Subsequently, the focus shifts to deep learning, in which the fundamental principles and specific techniques relevant to deep learning -based sleep analysis are presented. In addition, the key performance metrics used to evaluate the performance and reliability of deep learning -based sleep analysis methods are reviewed. Finally, the chapter concludes with a brief review of the current state of deep learning in sleep analysis.

2.1 Sleep architecture

Conventionally, sleep is measured using electroencephalogram (EEG) recordings to detect the electrical activity of the brain, accompanied by the simultaneous measurement of eye movements (EOG) and chin muscle tone (EMG). Based on the EEG signal frequencies, features and patterns detected in these signals, sleep can be categorized into five different stages according to the clinical scoring manual provided by the American Academy of Sleep Medicine (AASM) [17]. These stages include wakefulness (W), rapid eye movement (REM) sleep (R), and three stages of non-REM sleep (N1, N2, and N3). Each sleep stage facilitates different functions important for overall health and performance [1]. During clinical evaluation of sleep, the collected data is segmented into 30-second long, non-overlapping sequential epochs and a sleep stage is assigned to each epoch [17]. If characteristics of more than one sleep stage exist in one epoch, the sleep stage is determined based on the sleep stage most of the epoch consists of [17].

The W-stage includes the range from full alertness to the early stages of drowsiness until the onset of sleep. For most individuals a sinusoidal-like 8-13 Hz alpha rhythm recorded from the occipital region is present in the W-stage when their eyes are closed. Eye movements, such as eye blinks at the frequency of 0.5-2 Hz, eye movements indicating reading, or irregular rapid eye movements associated with a normal or elevated muscle tone in the chin are markers for the wakefulness stage. If any of these eye movements are

present in the epoch, or if more than 50 % of the epoch includes alpha rhythms, the epoch is scored as W [17]. An example of the EEG activity and the eye movements tracked with the electro-oculogram (EOG) in stage W is presented in Figure 1A.

The N1 stage typically marks the transition from wakefulness to sleep (sleep onset), in which the body shifts from a state of full alertness to lighter stages of sleep. In N1 sleep (Figure 1B), the brain waves shift from alpha waves observed in relaxed wakefulness to low-amplitude (4-7 Hz) activity indicating the onset of sleep [17]. Short-duration sharply contoured vertex sharp waves may be distinguished from the background signal. Conjugate and regular slow eye movements may also be observed [17]. Only about 5 % of the total sleep time is usually N1 sleep [18].

The EEG activity of N2 sleep (Figure 1C) is characterized by occasional k-complexes, which consist of a sharp negative wave immediately followed by a positive component, and by short bursts of distinct 11-16 Hz brain activity known as sleep spindles [17]. Eye movements usually cease in N2 sleep, although they may persist in some subjects. N2 is scored if the first half of the scored epoch or the latter half of the previous epoch exhibits these characteristic brain activities [17]. The following epochs are scored as N2 even in the absence of k-complexes or sleep spindles, until indicators of a transition to another sleep stage are observed [17]. Typically, the majority of sleep is spent in N2, which accounts for approximately 50 % of total sleep time [18].

In N3 sleep (Figure 1D), high-amplitude, slow 0.5-2 Hz waves dominate the EEG-signal [17]. If 20 % or more of the epoch contains slow wave activity, the epoch is scored as N3 even though some characterizing features of N2 sleep, particularly sleep spindles, may persist in N3 sleep [17]. Eye movements are typically not observed in N3 sleep [17]. Approximately 20 % of sleep is spent in N3 sleep [18].

The R-stage (Figure 1E) is often known as the “dreaming stage”, although some form of dreaming can also occur in nREM sleep [1]. True to its name, REM sleep is characterized by occasional rapid, sharp eye movements accompanied with a low chin muscle tone baseline and low-amplitude mixed-frequency EEG activity resembling that seen in stage N1, with possible trains of sharp 2-6 Hz sawtooth-like waves [17]. Short, irregular bursts of transient muscle activity may occur in REM sleep [17]. Scoring the REM stages requires assessing information from multiple surrounding epochs [17]. REM sleep usually accounts for 20-25 % of the total sleep time [18].

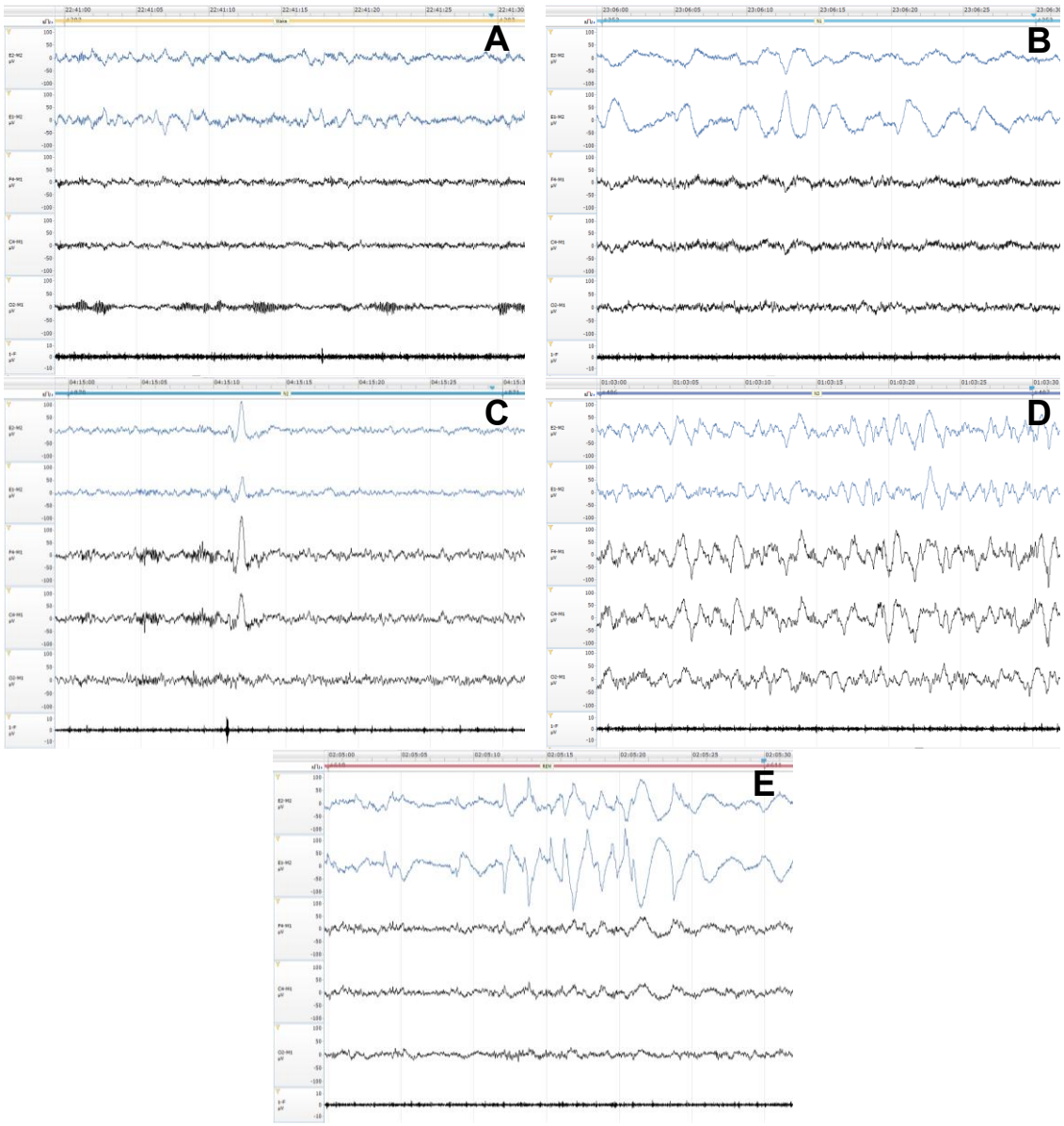

Figure 1. *Examples of measured signals in different sleep stages in the following order from top to bottom: EOG channels E1-M2 and E2-M2, EEG channels F4-M1, C4-M1, and O2-M1 and one submental EMG channel. The presented sleep stages are (A) Wake (stage W) (B) N1 sleep (C) N2 Sleep (D) N3 Sleep and (E) REM sleep (stage R). The length of the signal windows is approximately 30 seconds.*

Getting 7-9 hours of sleep over a single night is considered a normal amount of sleep [1]. Typically, sleep stages follow each other in approximately 90-minute long cycles throughout the night starting from the onset of lighter sleep followed by a gradual transition towards deeper sleep and eventually REM sleep [1]. However, the relative proportion of each sleep type in one cycle can vary throughout the night. For example, usually more N3 sleep occurs in early in the night, whereas the amount of REM sleep increases later towards morning [18]. Following sleep deprivation, the amount of N3 sleep increases, it occurs faster, and the measured signal amplitude may even be increased [1]. The loss

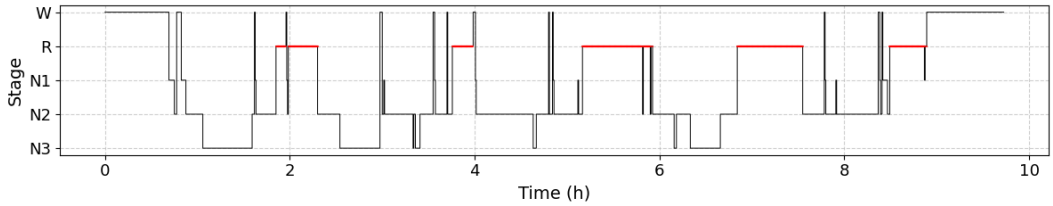

Figure 2. *An example hypnogram illustrating sleep architecture across a single night of sleep. Sleep is assessed in 30-second epochs, with a sleep stage assigned to each epoch based on the recorded physiological signals.*

of deep sleep is generally not as well tolerated as the loss of REM sleep [1]. Sleep architecture can be visualized using a hypnogram, where the sleep stages of consecutive epochs are plotted as a function of time (Figure 2).

Sleep stage scoring requires the consideration multiple parameters, often taking both previous and future epochs into account [17], making the process relatively complex and labor-intensive [5], [8]. Sleep stage characteristics can vary among individuals, and some features may overlap across different stages [17]. This difficulty can be further elevated in recordings from individuals suffering from sleep disorders. Owing to this complexity, it is challenging to create a scoring manual that accounts for all variations in sleep across the population [5]. Additionally, the manually conducted sleep stage evaluation inherently includes some subjectivity, which, combined with variability in the measured signals, is known to lead to some disagreement between sleep scorers. Several studies have examined the agreement between sleep scorers using various methodologies and metrics, with percent agreement varying in the range of 65–91 % and agreement measured with Cohen's kappa (see section 2.4.5) in the range of 0.38–0.94 [19], [20], [21], [22], [23], [24].

2.2 Obstructive Sleep Apnea

OSA is a common sleep-related breathing disorder that reduces sleep quality and is associated with the development of other health conditions including cardiovascular diseases, type 2 diabetes, cognitive impairment and mood disorders [3], [4]. The more immediate effects of OSA include excessive daytime sleepiness, fatigue, recurring awakenings from sleep, unrefreshing sleep, and impaired cognition [3], [4]. OSA is characterized by repeated restrictions in airflow to the lungs during sleep due to obstructions caused by soft tissues in the upper airways [3], [4]. Complete cessations of airflow are referred to as apneas, whereas partial limitations in the airflow are called hypopneas [17]. The adverse health effects resulting from recurring obstructions of airflow can be attributed to three key pathophysiological mechanisms: intermittent hypoxemia, sleep fragmentation and large intrathoracic pressure swings [4].

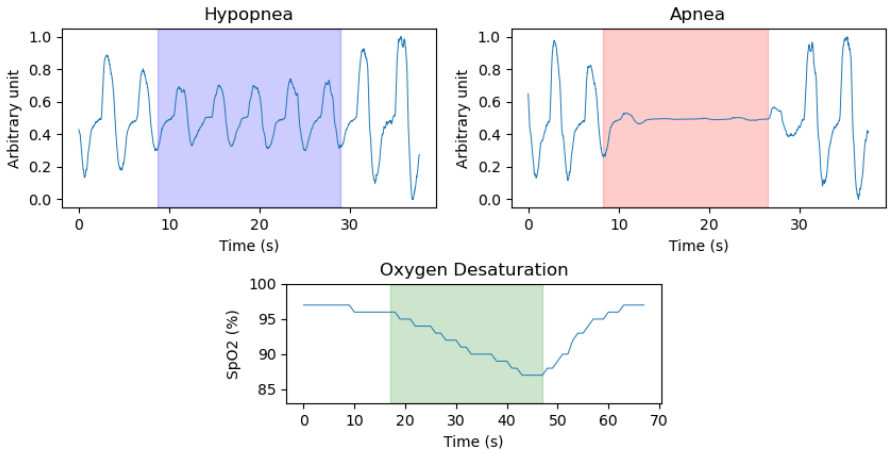

Figure 3. *The upper part of the figure presents examples of nasal pressure signals illustrating a hypopnea (highlighted in blue) and an apnea (highlighted in red). The lower part displays a desaturation event (highlighted in green) in the oxygen saturation signal.*

The American Association for Sleep Medicine (AASM) [17] defines apneas as events where measured airflow has dropped at least 90 % from the pre-event baseline for at least 10 s. Hypopneas are defined as events where the measured airflow drops at least 30 % from the pre-event baseline for at least 10 s, and the event is accompanied by either a 3 % drop from the baseline of the blood oxygen saturation levels or the event is associated with an arousal [17]. Figure 3 shows an example of an apnea, a hypopnea and an oxygen desaturation event. An arousal related to an apnea or hypopnea can help to restore the decreased muscle tonus, and lead to the reopening of the collapsed airway and restoration of the airflow through the upper airways [3], [4]. Although some arousals occur spontaneously without related pathophysiology, arousals related to apneas and hypopneas result in pathological sleep fragmentation, decreased sleep quality and contribute to the daytime symptoms [1], [3]. Additionally, OSA patients often have a decreased amount of N3 and REM sleep, accompanied by an increase in the amount of N1 and N2 sleep [1]. In particular, N3 sleep is closely associated with the restorative and refreshing qualities of sleep [1].

Risk factors for developing OSA include obesity, age, male sex, smoking and the use of alcohol [3]. The anatomical structure of the throat region, such as a narrow airway or enlarged tissues in the oropharyngeal area, can also contribute to the development of obstructive sleep apnea in an individual [3]. OSA is diagnosed based on OSA symptoms and existing comorbidities accompanied by an evaluation of the number of obstructive respiratory events during sleep determined from an overnight recording [25]. From this overnight recording, the apnea-hypopnea index (AHI) can be determined. AHI represents the number of apneas and hypopneas per hour of sleep and is calculated based on the total sleep time (TST) [1], [4]. An AHI of ≥ 5 is considered abnormal and serves as a threshold for OSA diagnosis, whereas if AHI is ≥ 15, the diagnosis can be reached

independently of daytime or nocturnal symptoms and existing comorbidities [26]. The current diagnostic standards include three OSA severity categories based on AHI: mild ($5 \leq AHI < 15$), moderate ($15 \leq AHI < 30$), and severe ($AHI \geq 30$) [1].

OSA can be treated by lifestyle changes such as weight loss or by using continuous positive airway pressure (CPAP) masks or mandibular advancement devices (MAD) during sleep to prevent obstruction of the airways [1]. If the occurrence of breathing cessations are sleeping-position dependent, preventing sleeping in that position may be used [4]. Surgical interventions are possible for select patients [4]. Successful treatment can reduce the adverse health effects related to OSA [4], which makes prompt and accurate diagnosis desirable at both the individual and societal levels.

2.3 Measuring sleep and breathing

Sleep studies have traditionally been classified into four categories ranging from Type I to Type IV [25]. Category I includes in-lab and category II home-based PSG measurements. A PSG measurement is considered the gold standard for OSA diagnostics, providing simultaneous monitoring of sleep, cardiac activity, and respiration. Up to 8 quantities are commonly measured in PSG: airflow, respiratory effort, oxygen saturation, snoring, electrical activity of the heart, body position, leg movement and sleep architecture [4]. These quantities are listed in Table 1, with a brief description of the measurement methods commonly used for the signals.

Sleep architecture is conventionally assessed using a combination of EEG, EOG, and submental EMG, as illustrated in Figure 4. EEG monitoring measures the electrical activity of the brain and involves three measurement electrodes placed on each side of the head, based on the recommendations by AASM [17]. Following the standard 10-20 system, these EEG derivations are F4-M1, C4-M1, and O2-M1 on the right side, with mirrored derivations (F3-M2, C3-M2, and O1-M2) on the left serving as backups in the case of electrode malfunction [17]. In addition to sleep staging, transient changes in the electrical activity of the brain, such as arousals, can be detected from the measured EEG signals and can be correlated with changes in respiratory-related parameters [3]. The EOG-electrodes used for tracking eye movements are placed 1 cm below the left outer canthus of the eye (E1), and 1 cm above the outer canthus of the right eye (E2), with both signals referenced to the M2 electrode [17]. Submental EMG is recorded using three electrodes placed near the inferior edge of the mandible: one at the midline 1 cm above the edge, and two positioned 2 cm below the edge, one to the left and one to the right of the midline [17].

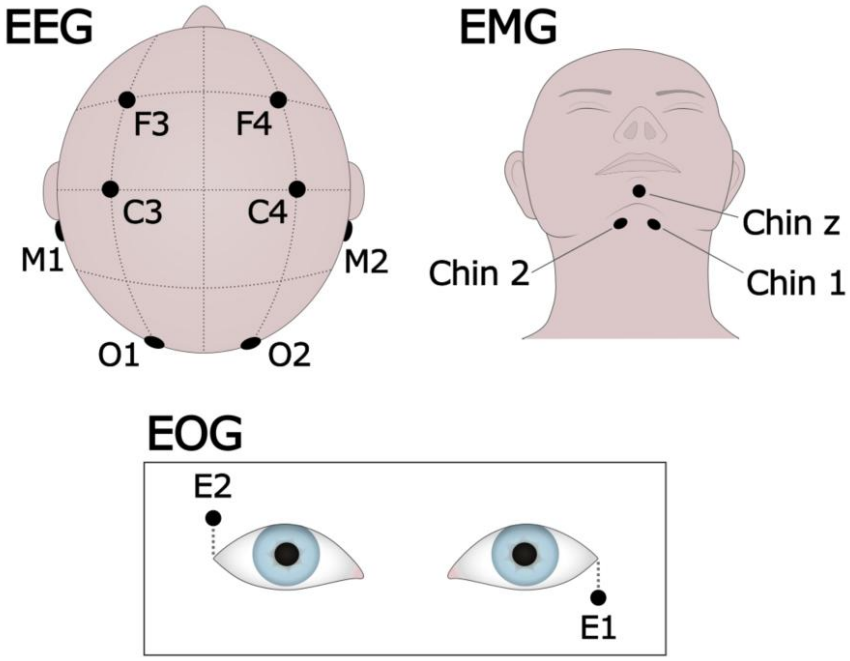

Figure 4. *Illustration of the placements of EEG, EMG and EOG electrodes for conventional sleep staging, as recommended by the American Academy of Sleep Medicine [17].*

Table 1. *Channels included in polysomnography measurements [1], [25].*

Measured quantity	Measurement methods	Purpose
Airflow	Nasal cannula connected to a pressure transducer *and/or* A thermal sensor placed on the path of nasal and oral airflow	Assessing the airflow through the upper airways
Respiratory effort	Thoracic and abdominal respiratory inductance bands *and/or* stomach and diaphragm EMG	Detecting respiratory effort and assessing the nature of breathing cessations
Oxygen saturation	Finger pulse oximeter	Assessing the blood oxygen saturation levels
Snoring	A microphone *and/or* Analyzing the high frequency components included in the nasal pressure measurement	Auditory assessment breathing and its restrictions during sleep
Sleep stages and arousals	EEG *and* EOG *and* submental EMG	Assessing sleep architecture
ECG findings	Two-electrode ECG	Detecting the electrical activity and arrythmias of the heart
Body position	Video (in-lab PSG) Positional sensor	Detecting sleeping position related OSA, aiding in assessment of sleep vs. wake
Leg movement	Leg EMG	Detecting abnormal movements during sleep

In the absence of suspicion of other sleep-related conditions, Type III home-based PG measurements limited to cardiorespiratory information may be used for OSA diagnosis and type IV home-based measurements with even further limited channels for OSA screening [1], [25]. Although PG is a more cost-effective and less invasive measurement, the absence of EEG monitoring prevents the assessment of sleep architecture and the accurate estimation of total sleep time. As AHI is defined based on total sleep time, in PG measurements an alternative metric, the respiratory event index (REI), is used [1]. REI defines the number of apneas and hypopneas evaluated over the recording time, as the wake portions of the recordings are not excluded [1]. Additionally, hypopneas leading to an arousal remain undetected with the PG measurement setup. These factors may decrease the number of detected hypopneas and can result in a decrease the estimated severity of OSA [13].

To address the issue related to PG measurements, machine learning approaches utilizing the photoplethysmogram (PPG) signal have been explored as an alternative method for assessing sleep [14], [15], [16]. PPG is a non-invasive method for assessing blood volume changes in the microvascular bed, commonly measured at peripheral sites such as the fingers [27]. The measurement is based on variations in absorbed (transmissive type) or reflected (reflective type) light corresponding to fluctuations in blood volume and concentrations of substances in the blood, allowing the measurement of physiological markers, such as heart rate and heart rate variability [28]. Additionally, by utilizing two different wavelengths of light in the measurement, typically red and infrared, blood oxygen saturation levels can be determined based on the differential absorption of these wavelengths by oxygenated and deoxygenated hemoglobin [28]. The principle of the transmissive-type PPG measurement used in this study, along with an example of the resulting PPG waveform, is illustrated in Figure 5. The changes in the amplitude and the frequency content of the PPG signal across sleep stages allow for the PPG-based sleep stage discrimination in automated analysis approaches [14].

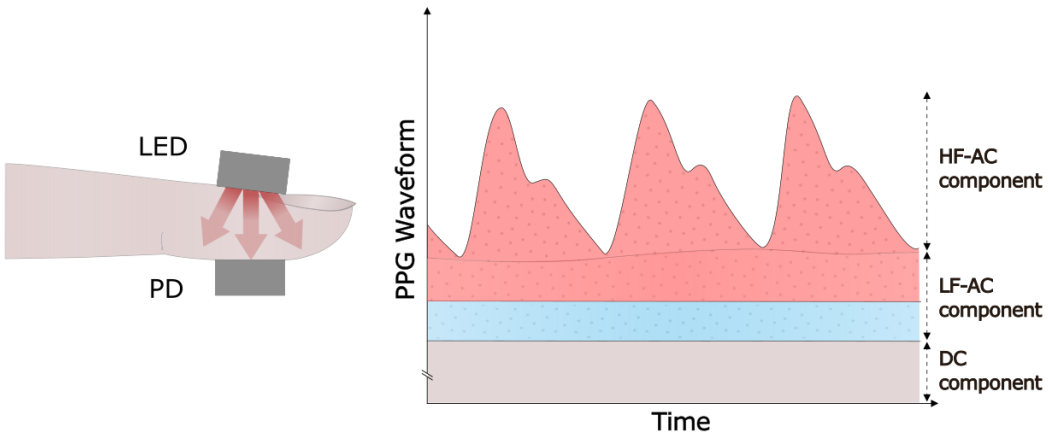

Figure 5. *The operating principle of transmissive photoplethysmography. A LED light source emits light trough tissues in a peripheral site, and a photodetector (PD) registers the amount of light transmitted through the tissue. The resulting PPG signal varies along with the cardiac activity (High Frequency Alternating Current component, HF-AC). The Low Frequency Alternating Current (LF-AC) represents the smaller changes in arterial and venous blood volume over time, affected by factors such as breathing, thermal regulation, autonomic nervous system activity, and hemoglobin concentration. The DC component presents the more constant absorption of the tissues. Figure adapted from [28].*

2.4 Deep learning

Deep learning is a subset of machine learning, a branch of artificial intelligence that enables systems to learn from data without being explicitly programmed to do so, enabling tasks such as classification, regression, and pattern recognition [29]. Deep learning utilizes artificial neurons arranged in neural networks inspired by biological neural systems [30]. A network of artificial neurons can be formed in different architectures by modifying how the artificial neurons are arranged and connected in the network [30]. One of the greatest advantages of deep learning compared to traditional machine learning approaches is that neural networks can learn from a raw data input removing the need for manual extraction and selection of the relevant features of the data [29]. Additionally, neural networks excel in processing large amount of data, making them well-suited for sleep studies [5].

Deep learning primarily encompasses two main learning paradigms: supervised learning and unsupervised learning, which are distinguished by the presence or absence of pre-labeled data during training [30]. Additionally, reinforcement learning is regarded as a third learning paradigm, wherein the learning occurs as the learning agent interacts with the environment and receives feedback [29]. In this study, the employed deep learning models for sleep staging and respiratory event recognition were trained using a supervised learning approach, where the sleep stages and events of the training data are manually pre-labeled. Therefore, the discussion in this section continues assuming this approach. In the following sections, the principles of neural networks are presented using

a standard fully connected feedforward neural network as a starting point for discussion to form a baseline understanding. After this, the following sections will further delve into convolutional neural networks (CNN) and recurrent neural networks (RNN), which are neural network architectures commonly used in automated analysis of sleep [5].

2.4.1 Feedforward Neural Networks

The goal of a deep learning network is to approximate complex functions by learning hierarchical representations of the input data through multiple layers of nonlinear transformations [31]. This can be mathematically understood by the expression

$$y = f(x; \theta) \tag{1}$$

where the neural network represented as f learns the values for the parameters θ that enable the best approximation for mapping the input x correctly to a prediction y [30].

In a neural network, neurons are organized into layers, with the connections between them determining how information flows through the network [30]. In a fully connected feedforward neural network, each artificial neuron of a single layer is connected to all the neurons in the previous and following layer, and the network processes the information unidirectionally [30]. The network includes three layer types: the input layer, hidden layer(s), and the output layer [30]. The input layer receives the raw data and feeds it forward into the hidden layers, with each neuron in the input layer representing a feature of the raw input [30]. For example, in image classification, each neuron in the input layer corresponds to a pixel value of the image [31], whereas in EEG data, a single neuron represents a single data point (sample) from the time-series signal [5]. As the data passes through one or more hidden layers, relevant features are extracted, and ultimately the processed information is received by the output layer providing the predicted classes or values of the neural network [30]. Figure 6 depicts a simple example of the structure of a feedforward multilayer neural network and its connections.

The connections between neurons are essentially mathematical relationships that can be broken down into two steps: the linear and non-linear phase [30]. First, a weighted sum of all the outputs from the neurons from the previous layer is calculated, and the bias value is added to the weighted sum [30]. For a single neuron i in layer l, this weighted sum z_i^l can be expressed as

$$z_i^l = \sum_j w_{i,j}^l x_j^{l-1} + b_i^l \tag{2}$$

where x_j^{l-1} is the output from neuron j in the previous layer, $w_{i,j}^l$ is the weight applied to the output from neuron j and b_i^l is the bias added to the weighted sum [30], [31].

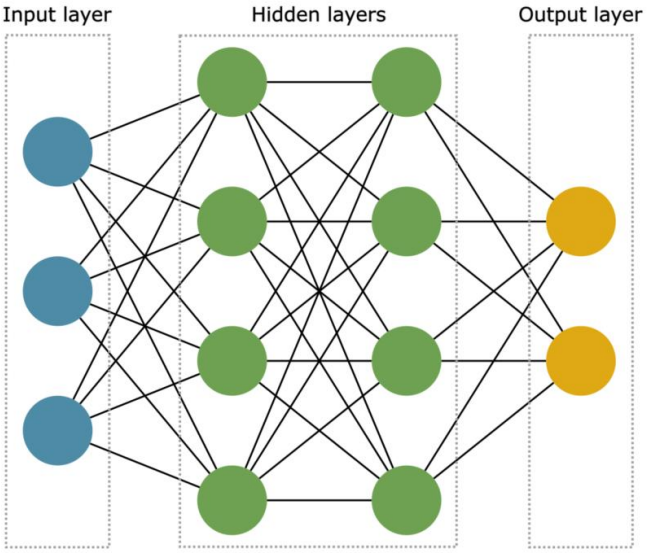

Figure 6. *An example representation of a fully connected neural network, with its artificial neurons arranged into layers and the neurons forming connections with each neuron in the preceding and following layer. The input layer receives the raw data and each of its neurons (depicted as coloured circles) represent a feature of the input data. The neurons in the hidden layers (two layers in this example) handle the transformations and feature extraction of the data. Eventually the output layer produces the output of the network. Each neuron in the output layer represents the predicted classes or values. The number of neurons in a layer and the connections can vary for a specific network. Adapted from [31].*

Secondly, the weighted sum is passed through an activation function, determining whether the neuron "activates" (passes an output) to the following layer and what this output value is [30]. For the same neuron i in layer l, the activation a_i^l can be expressed as [31]:

$$a_i^l = f(z_i^l). \tag{3}$$

Here, f is the activation function that introduces non-linearity into the network [30], [31]. Common activation functions include sigmoid functions $\left(f(x)_{sigm} = \frac{1}{1-e^{-x}}\right)$, rectified linear units $(f(x)_{ReLU} = \max(0, x))$ and hyperbolic tangents $\left(f(x)_{tanh} = \frac{e^x - e^{-x}}{e^x + e^{-x}}\right)$ [29].

2.4.2 Convolutional Neural Networks

Convolutional Neural Networks (CNN) are a type of feedforward neural network currently among the most popular deep learning techniques [29]. They are particularly well-suited for data with a grid-like structure, such as 1D time-series data, which consist of sequential samples taken at regular intervals, and 2D image data, represented as a grid of pixels [30]. CNNs process input data in smaller sequences, which allows them to detect localized patterns in the data, such as edges and textures [29]. Compared to the standard

feedforward neural networks, CNNs can be operated with a reduced number of parameters, which makes them computationally more efficient and reduces the likelihood of overfitting [29].

To learn features of the data, CNNs utilize convolutional layers and pooling layers [29]. In convolutional layers a convolutional filter (kernel) is applied step-by-step over the input [29]. A kernel is a grid of numbers or values called weights, which can be adjusted during the training process [29]. Unlike in a fully connected neural network, a single neuron in the convolutional layer is connected to only a limited number of neurons in the input layer [30]. Each neuron in the convolutional layer facilitates one convolution operation [30]. In each convolution operation, the current input sequence on which the kernel is applied to is multiplied by the weights of the kernel, and the multiplication results are summed to form one output point [30]. When this operation is repeated over the entire input, the outputs from these repeated convolution operations form a feature map [30]. Depending on the kernel, the feature map can highlight different aspects of the input [30]. Several different kernels can be applied over the same input, each forming a unique feature map [29]. In addition to the kernel weights, the information retained in the feature maps is influenced by input padding, which helps preserve edge details, and by the stride, which determines the step size of the filter application [29]. Non-linearity is also added to the network after each layer containing learnable weight biases [29].

Because multiple feature maps are usually generated, the input for the subsequent layers grows rapidly. This increase can be mitigated by the pooling layer, where the feature maps are downsampled into smaller feature maps while preserving the key features of the data [29]. Several pooling methods exist, and among the most popular are min, max, and global average pooling (GAP) [29]. Like in the convolution operation, the min and max pooling methods apply a sliding window across the data. Instead of performing a weighted sum, the min and max pooling methods extract either the minimum or maximum value within each window, respectively [29]. The GAP pooling method combines the whole input into a single average value [29].

After feature extraction, classification is achieved using one or more fully connected layers [5], [29]. The final layer of the classifier is commonly a softmax layer [30]. The softmax layer scales the output predictions for each class into a probability distribution, from which the class with the highest probability is assigned as the final prediction [5].

2.4.3 Recurrent Neural Networks

Recurrent Neural Networks (RNN) are neural network architectures for processing sequential data [31]. RNNs have been successfully used in several applications, including

natural language processing, speech processing, and time-series predictions [31]. RNNs include feedback loops in their structure, which allow them to consider the current input together with previously received inputs [30]. In practice, in information from one or more previous time steps is presented as additional input to the current layer [30]. As RNNs can consider contextual information, they can process information from the sequences of multiple sleep epochs and multiple sleep stages, which is an important feature in sleep staging [5]. Several RNN architectures exist, with different ways to employ the recurrent connections [30]. Some RNN architectures utilized in sleep analysis include long-short-term memory networks and gated recurrent units [8].

2.4.4 Training deep learning models

Training a neural network refers to a process where the networks “learn” by adjusting its parameters until a desired output has been reached. Typically, this process involves a training dataset for which a training error is calculated; this is, an error measure between the target output and actual output of the network that is minimized during training [30]. For feedforward neural networks, this error is commonly quantified using a cost function that measures how far off the model's predictions are from the true values [30]. Through iterative backpropagation, the model adjusts its weights and biases using gradient-based learning, gradually minimizing the cost function to improve performance [30]. However, the aim of training is not to optimize the neural network on the dataset it has been trained on, but to have it perform well on similar inputs it has not encountered before [30]. This is referred to as the generalization of the model [30]. To assess generalization, model performance is usually evaluated in a test dataset that has not been used in the training process, and from which a generalization error (test error) can be obtained [30]. Training and generalization errors contribute to two main challenges in machine learning: the underfitting and overfitting of data [30]. Underfitting refers to a situation where the test error was not sufficiently minimized for the model to adequately learn relevant patterns in the training data [30]. Overfitting occurs when a model becomes too specialized in the training data, capturing noise or irrelevant patterns, which results in poor generalization to new, unseen data [30]. Among the most common methods for avoiding overfitting is early stopping, which uses a separate validation dataset to monitor the model’s performance and stops training when this performance no longer improves [30]. Some other methods to avoid overfitting and improve generalization during the training process include using simpler models [30] randomly dropping neurons or weights out of the training phase, artificially expanding the training data, or utilizing batch normalization to normalize the output at each layer [29].

In addition to the training process, the training data itself has a profound impact on the generalization of the model. For the model to perform well on unseen data, the training data must be representative of the problem being solved to allow the model to learn relevant patterns [30]. For example, in sleep staging, the training data must include a large enough number of sleep stages that capture the variations in sleep throughout the night and across the population. However, in most real-world applications, achieving a completely error-free model is unrealistic even with the best training paradigms [30]. Some problems are inherently stochastic, that is, they involve randomness that prevents perfect prediction [30]. Additionally, the training data available to the model is always limited, making it impossible to capture every input scenario that can influence the output [30]. In tasks such as sleep staging, human scorers do not always fully agree with each other [19], [20]. Because the training data in supervised learning approaches for sleep staging is labeled by humans, inherent subjectivity and potential labeling inconsistencies introduce additional variability into the input-output relationship.

2.4.5 Assessing the performance of neural networks

An integral part of the development of a deep learning model is to determine the performance goals for the model and choosing the metrics to assess if those goals are reached in the current setup [30]. Although achieving a completely error-free model is unrealistic in most real-world applications, selecting appropriate metrics and performance goals can enhance a model's suitability for a specific task [30]. Simply assessing the model's overall accuracy, the fraction of correct predictions from all predictions, may not be sufficient to meaningfully assess the performance of the model [30]. This is especially true when a class imbalance exists: if one class holds most of the results, a model can achieve high accuracy by simply classifying everything into that majority class [30]. In some cases, certain types of errors carry higher costs than others [30]. For example, in sleep staging, misclassifying a sleep stage as wake can be more problematic than confusing one sleep stage for another [5].

Besides accuracy, some other parameters for assessing the performance of the models include Cohen's kappa, sensitivity, specificity, precision, and F-score [32]. To calculate these parameters, occurrences of each possible combination of agreement and disagreement between the reference and automated analysis, are counted [30], [32]. The number of possible combinations depends on the number of classes. For a binary (yes/no type) classifier, there are only four possible combinations: two where the reference and model arrived at the same classification and two where they disagreed [32]. When multiple classes are involved, the number of possible combinations increases [32].

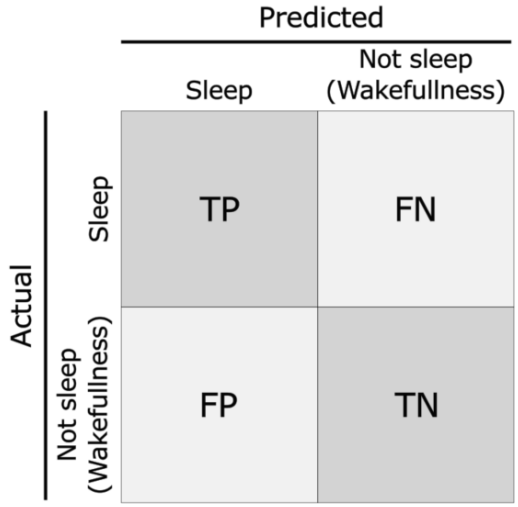

Figure 7. *An illustrative example of a confusion matrix for a two-class classification: sleep vs. "not sleep." The matrix demonstrates the agreement between predicted and reference labels. True positives (TP) represent correctly identified sleep epochs, while true negatives (TN) correspond to correctly identified "not sleep" epochs. False negatives (FN) occur when sleep epochs are misclassified as "not sleep," and false positives (FP) occur when "not sleep" epochs are misclassified as sleep. The term "not sleep" is used in this example instead of wakefulness to enhance clarity in terminology. Constructed based on [32] (see also supplementary material from [32]).*

For example, with five sleep stages, there are 25 possible combinations. The occurrences of these combinations are typically summarized in a confusion matrix [32]. An example of a two-class confusion matrix for predicting sleep and "not sleep" is shown in Figure 7, which illustrates the related concepts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Cohen's kappa is a metric used for assessing the overall agreement between the two raters classifying the same number of items into mutually exclusive categories while considering the possibility of agreement by chance [32]. It can be expressed as

$$\kappa = \frac{\mathrm{Acc} - p_e}{1 - p_e} \tag{4}$$

where Acc is the classification accuracy (the observed agreement) and p_e is the probability of chance agreement calculated as

$$p_e = \frac{1}{n^2} \sum_{g=1}^{G} n_{g1} n_{g2}, \tag{5}$$

where n is the total number of observations, G is the number of classes, n_{g1} is the number of observations of the first rater classified in class g, n_{g2} is the number of observations of the second rater classified in class g [32]. When the raters have perfect agreement, Cohen's kappa is 1, and a value of 0 indicates the level of agreement that would be expected purely by chance [33].

Sensitivity is expressed as the ratio of the number of true positives divided by the sum of true positives and false negatives [29], [32]:

$$sensitivity = \frac{TP}{TP+FN}. \quad (6)$$

True positives are samples correctly predicted by the model, whereas false negatives include the rest of the samples of the examined class that were misclassified [29]. A sensitivity of 1 indicates that the model classified correctly all occurrences of the examined class [32].

Specificity gives an understanding of how well the classifier rejects samples of other classes for a given class, expressed as the ratio of the true negatives and the sum of true negatives and false positives [29], [32]:

$$specificity = \frac{TN}{TN+FP}. \quad (7)$$

True negatives, like true positives, are samples where the model and the reference agree on the classification [29]. The difference can be best understood when considering true positives as the samples where the model and the reference agreed that the sample belongs to the examined class, and true negatives as instances where they agree that the occurrence does not belong to the examined class.

Precision p gives an understanding of how much of the model's predictions for a given class were correct and can be expressed as [29]:

$$p = \frac{TP}{TP+FP}. \quad (8)$$

Here FP is the number of false positives, the cases where the model classified a sample into the examined class when it actually belonged to another class [29]. A precision of 1 tells that all predictions by the model in the examined class are correctly classified [32].

Sensitivity and precision can be combined in the F-score, the harmonic average of these rates, given as [29]:

$$F = \frac{2pr}{p+r}. \quad (9)$$

Notably, precision and consequently also the F-score are affected by the prevalence of different classes, whereas sensitivity and specificity are not as they consider only the distributions of the examined class.

2.4.6 Deep learning in sleep analysis

Overnight PSG recordings produce several hours of data requiring labor-intensive and repetitive manual assessment, which has driven ongoing research into techniques to automate this process [5], [8]. Several different deep learning algorithms have been explored, including different types of feedforward and recurrent neural networks as well as combinations of these [5], [7], [8], [9], [10]. Different channels and channel combinations have also been used as inputs for deep learning models, including single or multiple EEG or EOG channels, various combinations of EEG and EOG channels, and, in some cases, one or more chin EMG channels are added to the input [5], [7], [8], [9], [10]. Rather than relying solely on traditional signals for sleep stage assessment, research has also investigated the possibility of extracting sleep information from unconventional signals. Most of these approaches rely on PPG signals [14], [16], and are sometimes combined with respiratory signals to enhance performance [15]. Additionally, simultaneous sleep staging and apnea and hypopnea detection has been conducted using different combinations of signals including PPG or EEG- and EOG channels [11].

Due to the complexity of sleep staging, complete agreement between manual and automated analysis is not expected [5]. Reaching an agreement of around 80 %, reflecting that of the inter-rater agreement of manual scorers, can be considered an excellent result for automated sleep staging analyses, while higher agreements may possibly indicate overfitting especially when the test data and training data are derived from the same dataset [5]. Agreements of this level have been reached in several studies using test data derived from the same datasets as the training data, and in a few cases, also using independent test datasets [5], [7], [8], [9], [10], [11].

3. MATERIALS AND METHODS

3.1 Clinical dataset and recording devices

The clinical PSG recordings (n = 342) analyzed in this study were recorded between 2019-2023 at Seinäjoki Central Hospital. This dataset consisted of PSG recordings from patients with suspected OSA or other sleep-related pathologies. Research and data permits (HYVAEP/424/2024/204994 and 340581) were obtained from the Wellbeing Services County of South-Ostrobothnia. During the data retrieval and pre-processing phase, 6 of the recordings were discarded because of technical issues in retrieving the recordings (lost recordings or lost clinical scoring information). Four recordings were excluded due to missing or practically unusable pulse oximeter data, and one due to technical issues. In total, 331 recordings were analyzed, including 193 females and 138 males. Of all recordings, 134 (44 males, 33 %) were classified as non-OSA based on the AHI score, while 112 (50 males, 45 %) recordings indicated mild OSA, 57 (27 males, 47 %) moderate OSA, and 28 (17 males, 61 %) severe OSA.

Table 2. *Demographic information of the clinical dataset (n = 331)*

	Median	**IQR**
Age (years)	37.0	26.0–50.0
Body mass index (kg/m^3)	28.7	24.9–33.2
AHI (1/h)	6.9	2.8–15.6
AHI Males (1/h)	7.9	3.4–17.7
AHI Females (1/h)	5.6	2.5–12.5
Arousal index (arousals/hour)	12.5	9.2–19.2
Total analysis time (min)	501.5	452.5–565.8
Total sleep time (min)	437.0	384.8–490.8
Wake after sleep onset (min)	32.0	15.5–65.8
Sleep latency (min)	16.5	6.5–33.3
N1 (%)	8.8	6.0–12.3
N2 (%)	55.5	49.9–61.5
N3 (%)	16.4	10.2–21.9
REM (%)	18.2	13.7–22.1
NREM (%)	81.8	77.9–86.3
Sleep efficiency (%)	87.8	80.7–86.3

Data recordings were conducted as ambulatory home recordings using commercially available Nox A1 and Nox A1S devices (Nox Medical Global, Reykjavik, Iceland). Qualified medical personnel performed the measurements. A total of 6 EEG channels (F3, C3, O1 with respect to M2 and F4, C4, and O2 with respect to M1, as presented in Figure 4) were measured with an additional reference electrode placed on the forehead. EOG electrodes E1 and E2 measuring eye movements were placed in the lower left corner

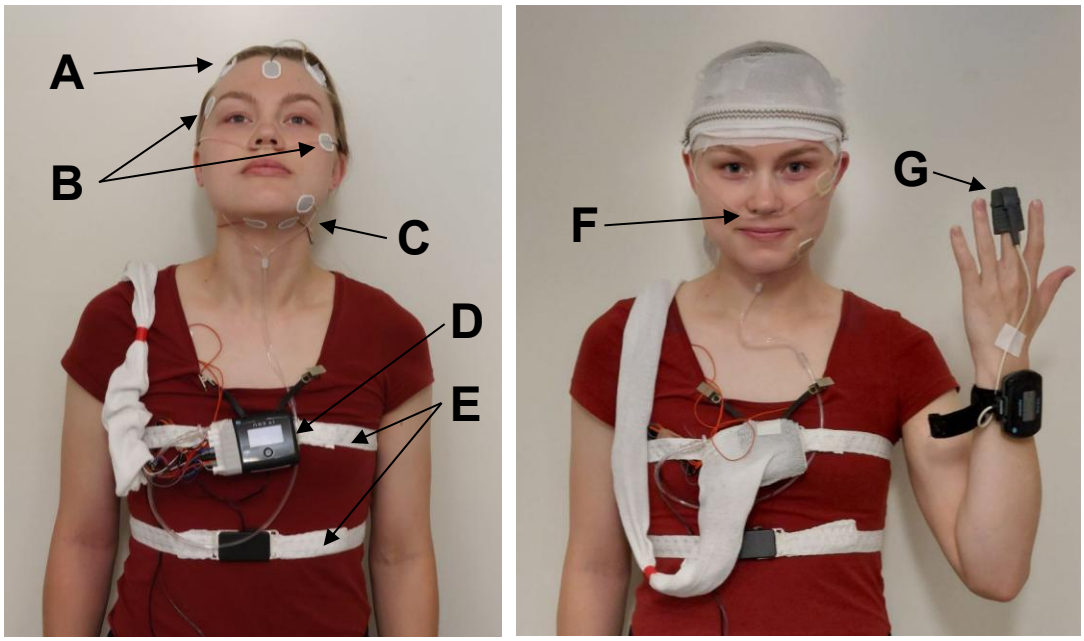

Figure 8. *The PSG measurement setup including (A) EEG-electrodes (all electrodes not visible), (B) EOG electrodes, (C) submental EMG electrodes, (D) a recording device with built-in audio and positional sensors (E) RIP belts (F) a nasal cannula and (G) a pulse oximeter. The ECG and leg EMG electrodes are not visible in these images.*

and the upper right corner of the eyes, respectively. Both EEG and EOG were measured at a sampling rate (f_s) of 200 Hz. Adhesive electrodes were used for electrode attachment when applicable. For regions with hair that prevented the attachment of adhesive electrodes, cup electrodes filled with conductive paste were used and attached with self-adhesive low-impedance cream. A mesh hood was placed on the head to further secure the electrodes in place. ECG was measured using two adhesive electrodes placed on the upper right region of the chest and the left side of the thorax with an f_s of 200 Hz. Submental EMG and EMG of the lower limbs were also measured using adhesive electrodes with a f_s of 200 Hz. Expansions of the thorax and abdomen were measured with two respiratory inductance plethysmography (RIP) belts with an f_s of 25 Hz, and nasal pressure was tracked with a nasal cannula (f_s = 200 Hz). PPG (f_s = 75 Hz) and SpO2 (f_s = 1 Hz for A1S devices and f_s = 3 Hz for A1 devices) were recorded from the patient's finger using a pulse oximeter. The sleeping position tracking (f_s = 20 Hz) was integrated into the recording device placed on the patient's chest. An example of the measurement setup is shown in Figure 8.

The recording window for each patient was programmed according to the patient's estimation of their sleep schedule in an effort to include the entire sleep time, which in most recordings resulted in some excess data before and after the onset and offset of sleep. This excess data was excluded from the analysis window, and the remaining data was

manually analyzed for sleep architecture and respiratory-related events by an experienced medical doctor specialized in sleep medicine following the AASM scoring rules. The analysis was conducted using Noxturnal software available for the clinician at the time of each recording. However, information regarding the exact software version(s) employed for scoring was not available. The sleep stage scores from this manual analysis served as a reference for the sleep architecture analysis performed by the deep learning model in this study.

3.2 Data pre-processing

The recordings and manual scorings were exported from Noxturnal 7.0.0. The exporting process was automated using STAR Sleep Recording Export software (SSRE) version 1.1.5. SSRE allows for some customization of the exported files [34], and is available for download at https://zenodo.org/records/11567841. In this study, a standard data export operation with the SSRE software generates four files per subject, stored in a subject-specific folder: an European Data Format (EDF) file containing the recorded signal data from the recording window, a text file containing recording and patient details and two XLS files, one containing sleep stage scorings and the other containing both sleep stage and event scorings from the analysis window.

After the exports, each subject folder was named with a unique uuid4-identifier. The data was then converted to Sleeplab-format using a python-based data format converter custom-made specifically for the conversion of this dataset. Sleeplab-format is a machine and human readable data format that can be used for storing and processing of PSG and PG data [35]. The data-format converter first reads the exported EDF files and stores each signal as NumPy-arrays and signal details as JSON files. It then reads the XLS file containing the manual scorings and scoring details, parses them and stores the sleep stages and respiratory-related events in separate JSON files. Finally, the converter reads the BMI, age, and sex of the patient from the TXT file and stores them in a JSON file. This conversion effectively pseudonymizes the data.

3.3 Computing environments for secondary-use data

According to the Finnish Act on the Secondary Use of Health and Social Data [36], the data provided under the data permit should be analyzed in a secure computing environment. According to these regulations, data analysis was conducted in a secure computing environment SD Desktop (CSC, Espoo, Finland) that is suitable for this purpose.

Following the instructions by CSC, a Findata-type project was first initialized, which allows the utilization of only a restricted version of the SD Desktop computing environment designed for handling sensitive secondary-use data. Within the project, an SD Desktop environment based on Ubuntu 22.04 was created for data analysis. As the SD Desktop environment includes only a limited set of default analysis tools, an Apptainer container was created in an Ubuntu 22.04-based virtual computing environment, launched via the cPouta cloud service (CSC, Espoo, Finland). The container includes Python 3.10.11 and relevant libraries for the deep learning model analysis, such as TensorFlow 2.12.0 and Keras 2.12.0 API. It was built using Docker as the bootstrap method, with Ubuntu 22.04 as the base image. A second container, created with the same instructions, was also created for the result analysis, containing Python libraries required for this task. The preprocessed dataset, the Apptainer containers, and Python scripts for analyses were securely transferred to the SD Desktop computing environment via SD Apply.

The results of this study produced using secondary-use data were scrutinized by the Finnish Social and Health Data Permit Authority Findata to ensure the anonymity of the results before accessing and publishing them outside the secure computing environment.

3.4 Deep learning tools and analysis

3.4.1 Deep learning models

Two trained deep learning models for simultaneous sleep stage and respiratory event classification were provided for this study by the Sleep Technology and Analytics Research (STAR) group. The first deep learning model is designed to be used with PSG data, taking as inputs an EEG signal from derivation F4-M1, an EOG signal from derivation E1-M2, and PPG waveform and blood oxygen saturation, nasal pressure, and RIP sum (the combination of abdominal and thoracic RIP belts) signals. The second deep learning model limits the signals to those typically available in PG measurements, incorporating pulse waveform, oxygen saturation, nasal pressure, and RIP sum signals in its analysis. The inclusion of EEG- and EOG-channels in the analysis is a natural choice as they are the basis for the conventional sleep staging [17]. Utilizing these channels have provided promising results in automated sleep staging [7], [9], [11]. The PPG signal has also demonstrated promise in sleep staging [11], [14], as it indirectly reflects the physiological state of the body. Additionally, the PPG measurement is both widely used and easy to apply, making it appealing for sleep staging. The remaining signals were selected based on their estimated potential to enable the detection of respiratory-related events.

Both models employ a fully convolutional U-Time-based architecture for sleep staging, first introduced by Perslev et al. [37] and later refined by Huttunen et al. [11]. This architecture comprises an encoder, a decoder, and two classifiers. The encoder processes the raw input data through six convolutional blocks (CB), each incorporating convolutional layers with batch normalization and max pooling. The output of the encoder is then processed using a technique called Atrous Spatial Pyramidic Pooling to capture multi-scale information [38]. The decoder continues the processing and feature extraction through five upsampling blocks (UBs). Each UB upsamples its input before passing it through convolutional and batch normalization layers. To preserve low-level features, a skip connection is used to concatenate feature maps from each EB with the corresponding UB of the same scale. Additionally, a Squeeze-and-Excitation module [39] is applied at the output of each CB and UB to enhance the model's generalizability and performance. Finally, the processed feature maps are passed through two parallel classifier branches: one for sleep staging and another for respiratory event detection. Each classifier consists of fully connected layers that refine the extracted features and produce the final predictions via a softmax layer. For sleep stages, the softmax scores were provided for each 30 s segments and for respiratory events, for each 1 s segments. However, within the scope of this work, only the sleep staging results were further analyzed. The model definitions are available at https://github.com/rikuhuttunen/psg-simultscoring-models.

3.4.2 Training data

A type I PSG dataset measured at Kempenhaeghe Center for Sleep Medicine was used for training both models. This dataset includes participants with no suspicion of sleep or sleep-related disorders and participants with suspected OSA. Parameters further describing the training data are presented in Table 3.

Table 3. *Descriptive statistics for the training dataset, including medians, minimum and maximum values for age, BMI, and AHI, as well as the number of recordings and the percentage of male and female participants.*

Parameter	Median	Min	Max
Age (years)	53	18	84
BMI (kg/m^3)	27.17	16.02	139.75
AHI (1/h)	11.66	0	99.61
Total number of recordings	774 (Males 62.27 %, Females 37.73 %)		

Two recordings from the dataset were excluded from this study due to missing RIP signals. From the remaining dataset (n = 772), a test set (n = 77) and a validation set (n = 70) were randomly selected, leaving the remaining PSG recordings to form the training

dataset (n = 625). After training, the model performance was evaluated against the test dataset to ensure that it can be generalized on unseen data, the results presented in Table 4 and Figure 9. The overall accuracy and Cohen's kappa values for the PSG model was 84.26 % and 0.78 and for the PG model 74.18 % and 0.64, respectively. The sensitivity and specificity for detecting sleep was 97.15 % and 89.13 % for the PSG model and 97.71 % and 77.95 % for the PG model, respectively.

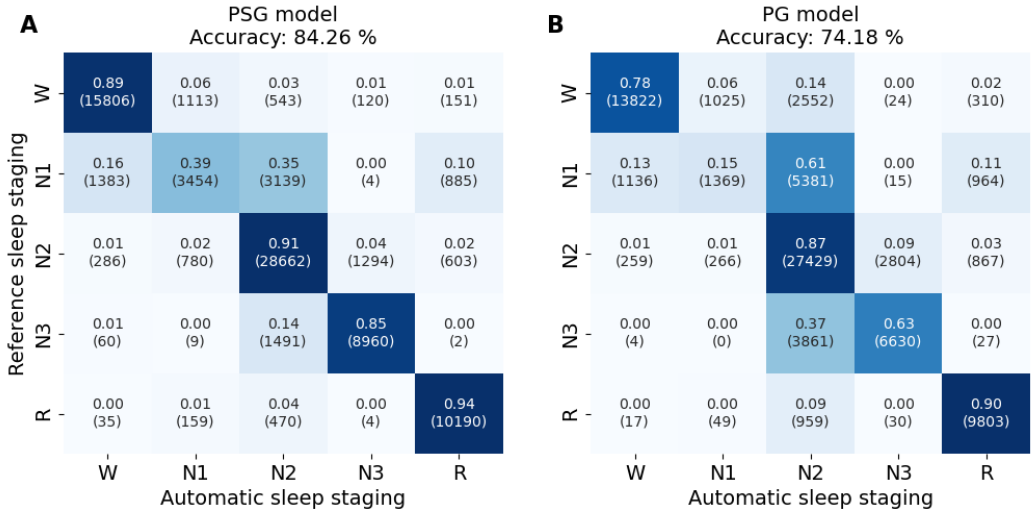

Figure 9. *Confusion matrices illustrating the performance of the two models evaluated on the test dataset.*

Table 4. *Performance metrics of the models evaluated on the test dataset.*

	PSG model			
	Precision	**Sensitivity**	**F1-score**	**Support**
W	0.90	0.89	0.89	17 733
N1	0.63	0.39	0.48	8 865
N2	0.83	0.91	0.87	31 625
N3	0.86	0.85	0.86	10 522
REM	0.86	0.94	0.90	10 858
Accuracy		84.26 %		
Kappa		0.78		
	PG model			
	Precision	**Sensitivity**	**F1-score**	**Support**
W	0.91	0.78	0.84	17 733
N1	0.50	0.15	0.24	8 865
N2	0.68	0.87	0.76	31 625
N3	0.70	0.63	0.66	10 522
REM	0.82	0.90	0.68	10 858
Accuracy		74.18 %		
Kappa		0.64		

3.4.3 Application of the deep learning models

A previously developed GUI interface was used to handle the analyses. To reduce computational load, the signals were linearly resampled to 64 Hz before analysis using NumPy's interp-function. For signals requiring downsampling, a low-pass filter was applied before resampling to avoid signal aliasing. As the entire recording was exported, the signals used for the analysis were trimmed to match the analysis window determined manually by the clinician. This ensured the 30-second sleep stage epochs were derived from the same segments of the signal as the manual analysis. Finally, the signals were z-score normalized. The data was then analyzed by the deep learning models in the Apptainer container's Python environment described previously in section 3.3, and the sleep staging results, including the relative probabilities for each stage per epoch, were stored in a JSON file.

3.4.4 Performance assessment

The sleep stage analysis by the two deep learning models was compared to the analysis conducted by the clinician on an epoch-by-epoch basis. The comparison was performed separately for the whole dataset, sex-based subgroups (males and females) and OSA severity-based subgroups (non-OSA, mild, moderate, and severe). Confusion matrices were formed from the comparison results.

The last sleep stage for each subject was discarded from the analysis. This was done to ensure that only the sleep stages within the analysis window intended by the clinician were included in the analysis. The analysis window in the Noxturnal software can be set arbitrarily at any time point of the recording. If the analysis window ends within an epoch, the entire 30-second epoch will still be exported. This means that even if only a small portion of an epoch was included, it would still be marked as a full 30-second epoch. However, the effect of this step on the overall results was observed to be negligible as the removed stages comprise only a minor portion of all analyzed sleep stages.

Overall accuracy and Cohen's kappa values were calculated for each of the analyzed groups, along with precision, sensitivity, and F1 scores for each sleep stage. These widely used metrics in automated sleep analysis provide a group-level estimation of model performance and help determine whether overall results align with the agreement observed between human scorers. Additionally, stage-specific metrics offer insight into how well individual sleep stages are detected and can reveal patterns that reflect or diverge from common misclassification trends seen in manual scoring. Sensitivity and specificity for differentiating sleep from wake were also computed for each group.

4. RESULTS

4.1 Performance evaluation on the entire dataset

Figure 10 presents the confusion matrices derived from the entire dataset for both the PSG and PG model. Both matrices have been row-normalized based on the reference analysis. The overall and per-class performance metrics (precision, sensitivity, and F1) are detailed in Table 5.

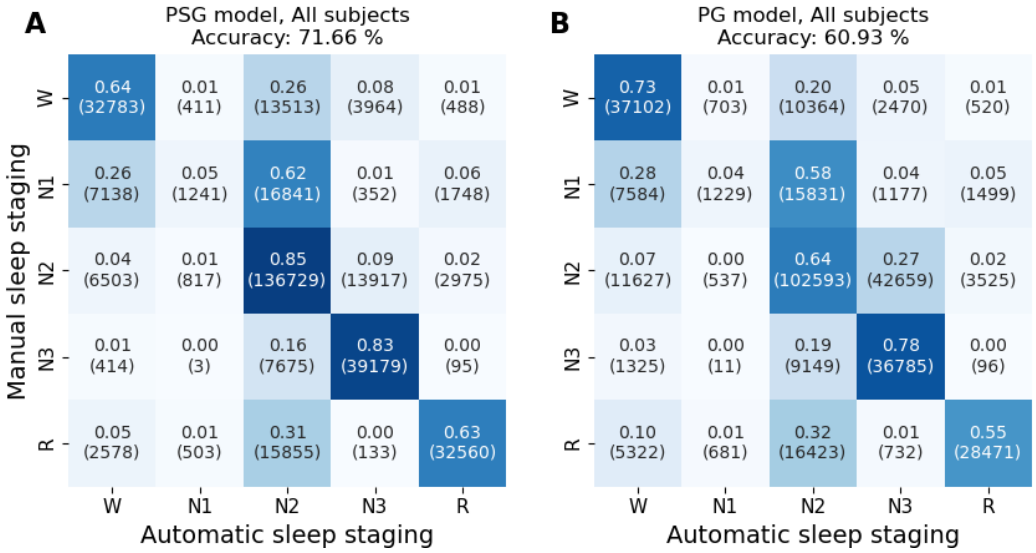

Figure 10. *Confusion matrices derived from the whole studied dataset using (A) the PSG model and (B) the PG model.*

Table 5. *Performance metrics of the automated analyses for the whole dataset, including per-class metrics precision, sensitivity and F1, the overall classification accuracy and agreement measured as Cohen's Kappa.*

		W	N1	N2	N3	REM
PSG model analysis	Precision	0.66	0.42	0.72	0.68	0.86
	Sensitivity	0.64	0.05	0.85	0.83	0.63
	F1	0.65	0.08	0.78	0.75	0.73
	Accuracy			71.66 %		
	Kappa			0.58		
PG model analysis	Precision	0.59	0.39	0.66	0.44	0.83
	Sensitivity	0.73	0.04	0.64	0.78	0.55
	F1	0.65	0.08	0.65	0.56	0.66
	Accuracy			60.93 %		
	Kappa			0.45		

The PSG model demonstrated the highest sensitivity for N2 and N3 sleep, whereas the PG model was most successful in correctly classifying N3 sleep and wake. However, both models exhibited poor sensitivity for N1 sleep. These epochs were most often misclassified as N2 sleep or wake. In general, both models most frequently misclassified

other sleep stages as N2 sleep, while N2 sleep itself was most often misclassified as N3 sleep. Notably, the PG model confused N2 with N3 even more frequently than the PSG model.

4.2 Sex-based performance evaluation

The row-normalized confusion matrices for the PSG and PG models analyses in males and females are presented in Figure 11.

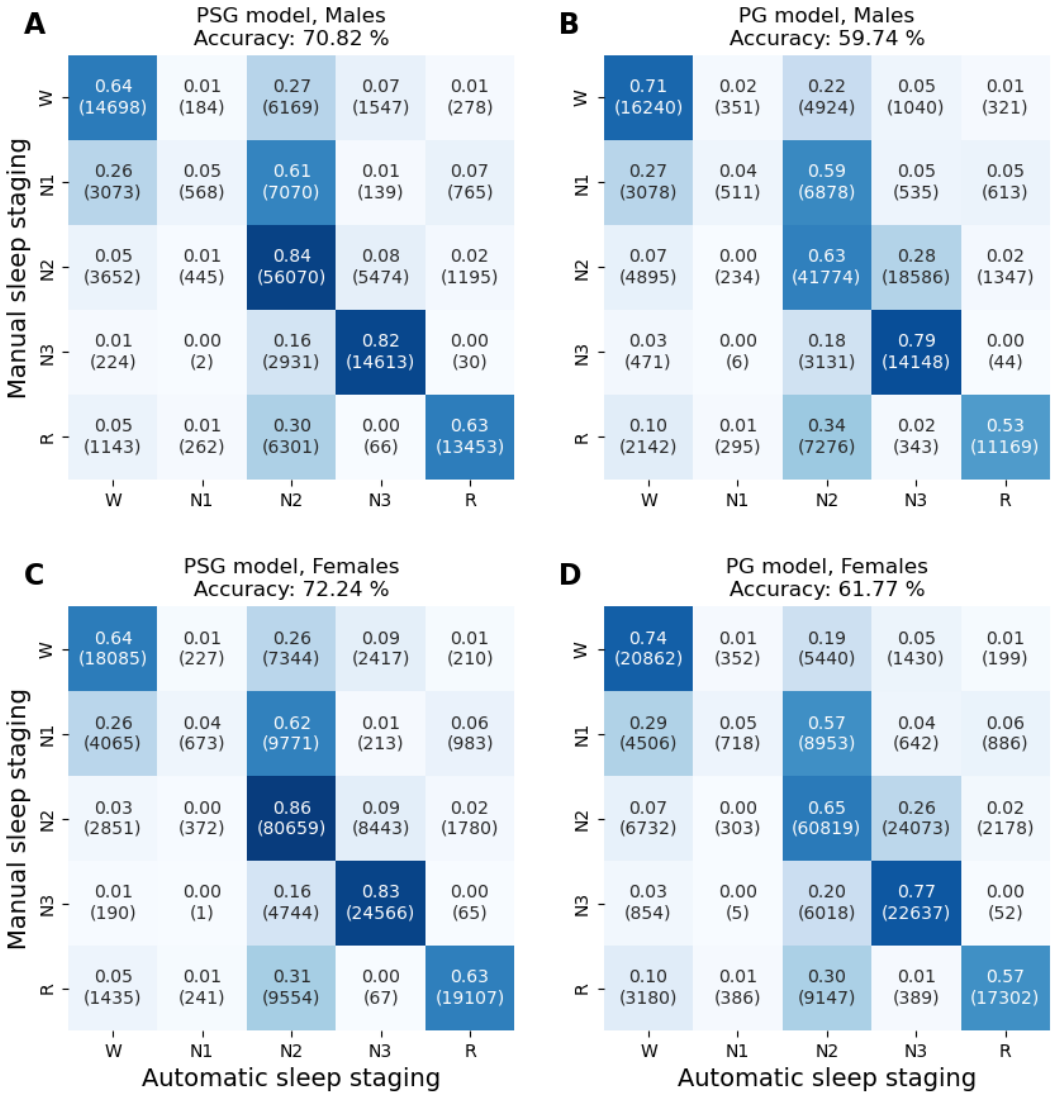

Figure 11. *Confusion matrices for sleep stage analysis derived from male and female participants in the dataset. (A) PSG model analysis for males. (B) PG model analysis for males. (C) PSG model analysis for females. (D) PG model analysis for females.*

Only subtle differences in the performance metrics were observed for the sex-based groups (Table 6). Both the classification accuracy and kappa values indicate that, overall, the model performed slightly better for females than for males.

Table 6. *Performance metrics of the automated analyses for sex-based groups, including per-class metrics precision, sensitivity and accuracy, the overall classification accuracy and agreement measured as Cohen's Kappa.*

PSG model analysis						
		W	**N1**	**N2**	**N3**	**REM**
Males	Precision	0.64	0.39	0.71	0.67	0.86
	Sensitivity	0.64	0.05	0.84	0.82	0.63
	F1	0.64	0.09	0.77	0.74	0.73
	Accuracy			70.82 %		
	Kappa			0.56		
Females	Precision	0.68	0.44	0.72	0.69	0.86
	Sensitivity	0.64	0.04	0.86	0.83	0.63
	F1	0.66	0.08	0.78	0.75	0.73
	Accuracy			72.24 %		
	Kappa			0.58		
PG model analysis						
		W	**N1**	**N2**	**N3**	**REM**
Males	Precision	0.61	0.37	0.65	0.41	0.83
	Sensitivity	0.71	0.04	0.63	0.79	0.53
	F1	0.65	0.08	0.64	0.54	0.64
	Accuracy			59.74 %		
	Kappa			0.43		
Females	Precision	0.58	0.41	0.67	0.46	0.84
	Sensitivity	0.74	0.05	0.65	0.77	0.57
	F1	0.65	0.08	0.66	0.58	0.68
	Accuracy			61.77 %		
	Kappa			0.46		

4.3 OSA Severity -based performance evaluation

4.3.1 PSG model analysis

The row-normalized confusion matrices for the PSG model analysis across OSA severity categories are presented in Figure 12. The analyses in the OSA severity groups showed a progressive decline in the percent accuracy and Cohen's kappa with increasing OSA severity. The decrease in both values was most pronounced between the moderate OSA group and the severe OSA group.

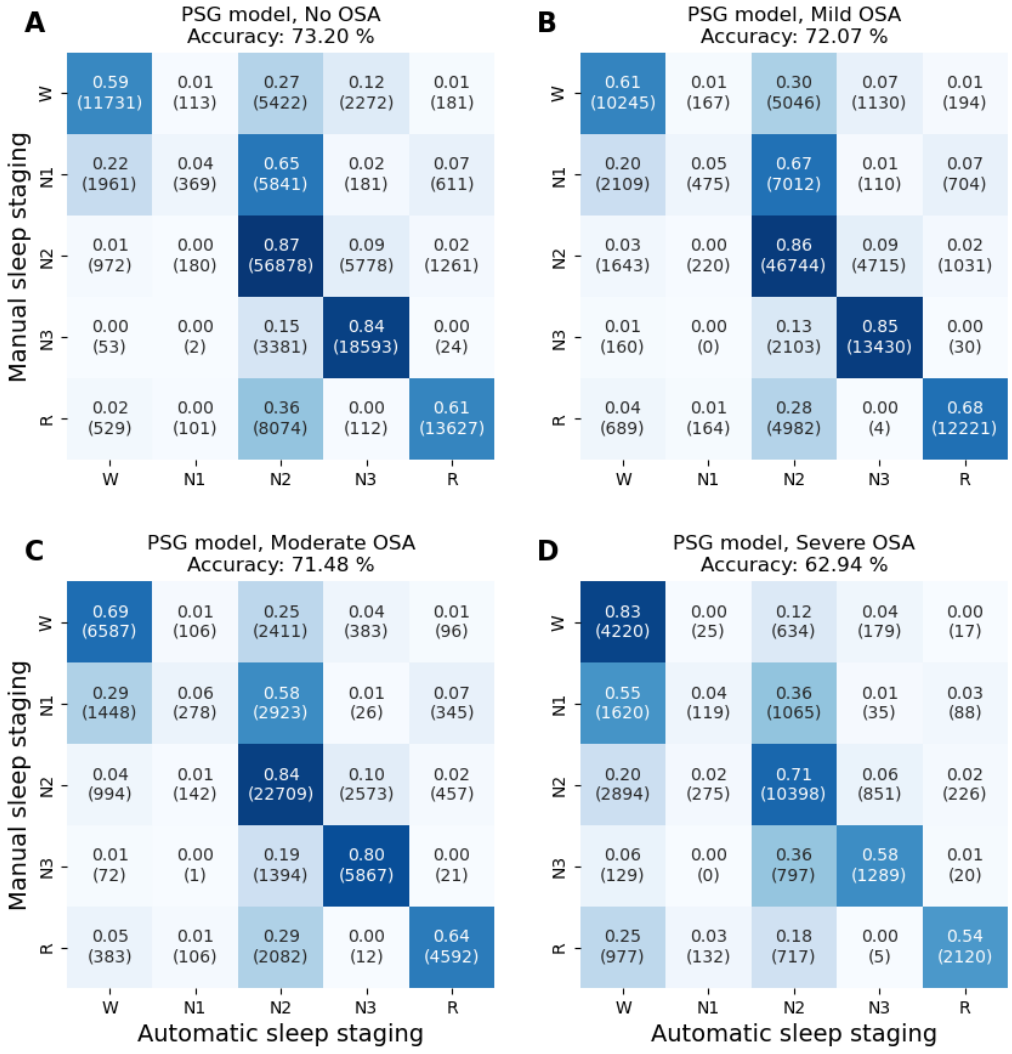

Figure 12. *Row-normalized confusion matrices from the PSG model, showing results for groups with (A) no recorded OSA, (B) mild OSA, (C) moderate OSA, and (D) severe OSA.*

When examining individual sleep stages, the performance is relatively similar between the no OSA, mild OSA and moderate OSA subgroups. In the severe OSA group, a shift

toward detecting wake was observed, with the model both identifying wake more accurately and misclassifying other stages as wake more frequently. The overall and per-class metrics are further detailed in Table 7.

Table 7. *Performance metrics of the PSG model analysis for the OSA severity subgroups, including per-class metrics precision, sensitivity and accuracy, the overall classification accuracy and agreement measured as Cohen's Kappa.*

PSG model analysis						
		W	**N1**	**N2**	**N3**	**REM**
No OSA	Precision	0.77	0.48	0.71	0.69	0.87
	Sensitivity	0.59	0.04	0.87	0.84	0.61
	F1	0.67	0.08	0.79	0.76	0.71
	Accuracy			73.20 %		
	Kappa			0.60		
Mild	Precision	0.69	0.46	0.71	0.69	0.86
	Sensitivity	0.61	0.05	0.86	0.85	0.68
	F1	0.65	0.08	0.78	0.76	0.76
	Accuracy			72.07 %		
	Kappa			0.58		
Moderate	Precision	0.69	0.43	0.72	0.66	0.83
	Sensitivity	0.69	0.06	0.84	0.80	0.64
	F1	0.69	0.10	0.78	0.72	0.72
	Accuracy			71.48 %		
	Kappa			0.57		
Severe	Precision	0.43	0.22	0.76	0.54	0.86
	Sensitivity	0.83	0.04	0.71	0.58	0.54
	F1	0.57	0.07	0.74	0.56	0.66
	Accuracy			62.94 %		
	Kappa			0.46		

4.3.2 PG model analysis

The PG model analyses for OSA severity groups showed a general decline in accuracy with increasing OSA severity, although the accuracy for the mild OSA group was slightly lower than for the moderate OSA group (Figure 13). Cohen's kappa values also declined across all severity groups.

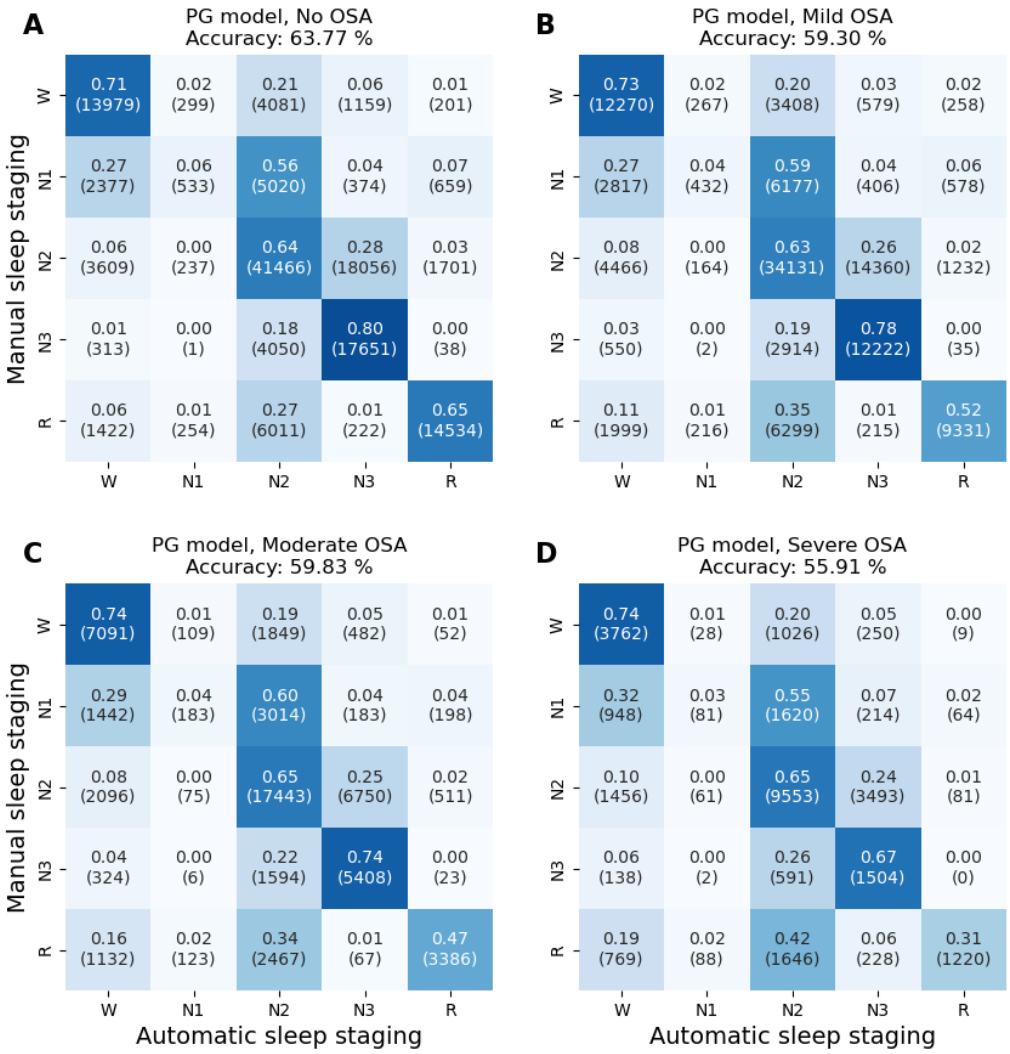

Figure 13. *Row-normalized confusion matrices from the PG models analysis, showing results for groups with (A) no recorded OSA, (B) mild OSA, (C) moderate OSA, and (D) severe OSA.*

For individual sleep stages, the proportion of correctly classified stages N3 and R decreased progressively with increasing OSA severity. This decline was particularly pronounced for stage R, where the sensitivity dropped from 0.65 in the no OSA group to 0.31 in the severe OSA group, mainly due to the increased frequency of confusing it as N2 sleep or wakefulness. The performance metrics for the OSA severity groups analyses are listed in Table 8.

Table 8. *Performance metrics of the PG model's analysis for the OSA severity subgroups, including per-class metrics precision, sensitivity and accuracy, the overall classification accuracy and agreement measured as Cohen's Kappa.*

PG model analysis						
		W	**N1**	**N2**	**N3**	**REM**
No OSA	Precision	0.64	0.40	0.68	0.47	0.85
	Sensitivity	0.71	0.06	0.64	0.80	0.65
	F1	0.68	0.10	0.66	0.59	0.73
	Accuracy			63.77 %		
	Kappa			0.49		
Mild	Precision	0.56	0.40	0.64	0.44	0.82
	Sensitivity	0.73	0.04	0.63	0.78	0.52
	F1	0.63	0.08	0.64	0.56	0.63
	Accuracy			59.30 %		
	Kappa			0.42		
Moderate	Precision	0.59	0.37	0.66	0.42	0.81
	Sensitivity	0.74	0.04	0.65	0.74	0.47
	F1	0.65	0.07	0.66	0.53	0.60
	Accuracy			59.83 %		
	Kappa			0.42		
Severe	Precision	0.53	0.31	0.66	0.26	0.89
	Sensitivity	0.74	0.03	0.65	0.67	0.31
	F1	0.62	0.05	0.66	0.38	0.46
	Accuracy			55.91 %		
	Kappa			0.35		

4.4 Detecting sleep and wake

For both models, the sensitivity and specificity for detecting sleep were evaluated across the entire dataset, as well as within sex-based and OSA severity-based subgroups (Figures 14 and 15).

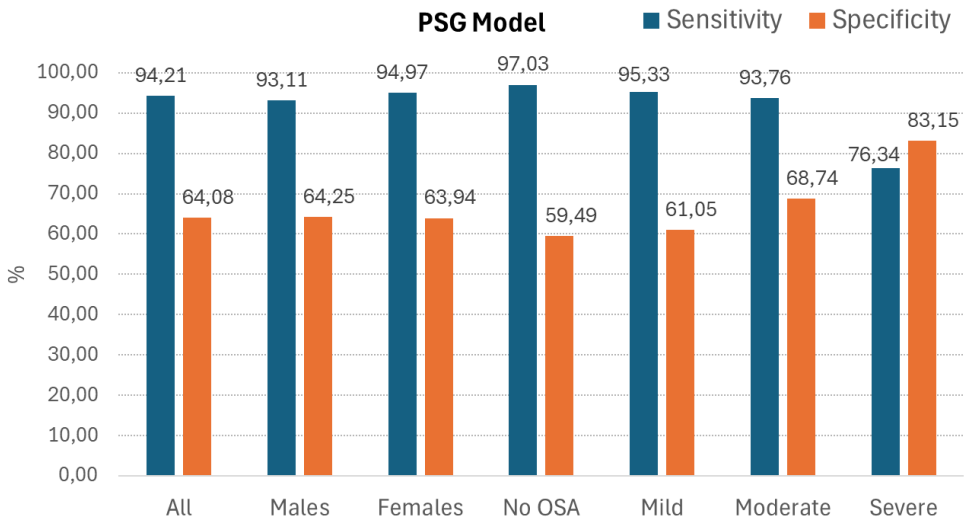

Figure 14. *Sensitivity and specificity for sleep detection across the entire dataset, sex-based groups (males and females), and OSA groups (no OSA, mild OSA, moderate OSA, severe OSA), calculated for the PSG model analyses.*

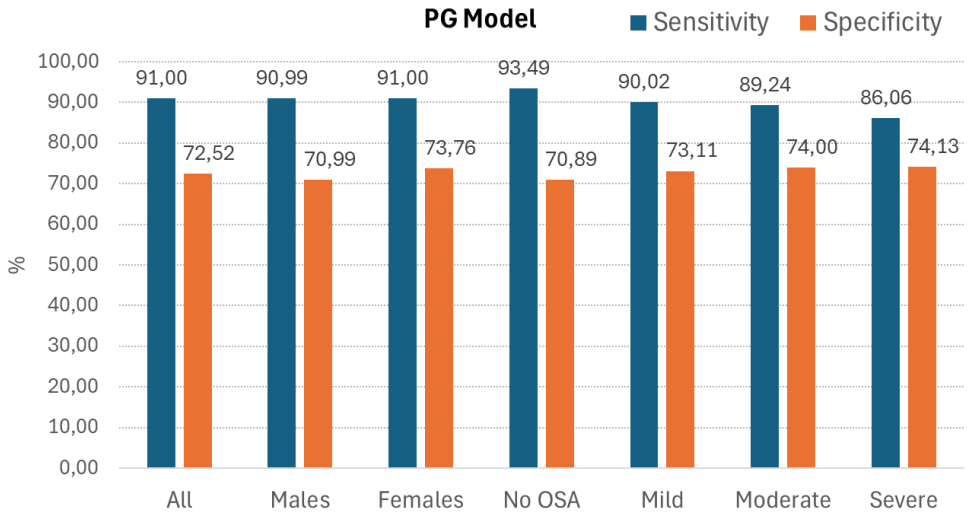

Figure 15. *Sensitivity and specificity for sleep detection across the entire dataset, sex-based groups (males and females), and OSA groups (no OSA, mild OSA, moderate OSA, severe OSA), calculated for the PG model analyses.*

5. DISCUSSION

In this study, two previously developed deep learning models for automated classification of sleep stages were tested on a clinical dataset (n=331). The first model is designed specifically for PSG data, including EEG (F4-M2), EOG (E1-M2), PPG, Nasal Pressure, SpO2, and RIP sum signals in its analysis. The second model was designed to be compatible with both PSG and PG data by limiting its input to PPG, Nasal Pressure, SpO2, and RIP sum signals, which are commonly available in PG recordings. To assess the performance of the automated analysis results were compared to the clinical evaluation by a clinician specialized in sleep medicine. In addition to analyzing the whole dataset, subgroup analyses were conducted to examine variations in analysis performance across sex-based and OSA severity-based groups.

The PSG and PG model analysis reached five-stage classification accuracies of 71.66 % and 60.39 %, respectively. The respective Cohen's Kappa values were 0.58 and 0.45, indicating moderate agreement between the manual and automated sleep staging according to Landis and Koch [40]. The difference in the two models performance aligns with the findings of Huttunen et al [11] who used a model with the same architecture as in this study while comparing input signal combinations for simultaneous sleep staging and respiratory event detection. They reported improved percent accuracy in sleep staging when an EEG channel was included in the signals used in the analysis (79 %) compared to two models that relied on PPG and SpO2 (69 %) or PPG, SpO2, and nasal pressure channels (70 %). A similar model architecture was also used in a study by Rusanen et al. [7], where the model performance in sleep staging was tested in clinical, open access, and self-applied PSG datasets, reaching accuracies of 80–83 % and kappa values of 0.71–0.75 using single EEG, single EOG, and EEG and EOG combinations. These results demonstrate that this type of model can, at least in some cases, be generalized beyond a test dataset derived from the same dataset as training data has been segmented from. For some other models combining CNN and RNN architectures, even higher agreement in sleep staging has been reported for analyses including EEG and EOG channels. For example, Korkalainen et al. [10] used a deep learning model combining a CNN and RNN, achieving overall accuracies of 83.9 % and Cohen's kappa values of 0.78 on a public dataset and 83.8 % and 0.78 on a clinical dataset using EEG (C4-M1) and EOG (E1-M2) channels. PPG-based analyses relying on CNN and RNN model architectures have provided accuracies of 64.1 % and kappa values of 0.51 [14] and 68.7 % and 0.60 [16] for five-stage sleep staging. Compared to these results from

other studies, both models used in this study demonstrated lower performance, with the results of the PSG model, which included an EEG and EOG channel in its analysis, being particularly reduced. However, it must be noted that the results of different studies using different training data, model architectures and test data are not directly comparable between each other.

A very small improvement in performance for both models, as indicated by the overall accuracy and Cohen's kappa values, was detected in females. However, the differences were subtle, and it seems reasonable to conclude that the model performed consistently in both sex-based subgroups. Moreover, the training data included more male than female subjects (62,27 %), whereas the clinical data had a greater proportion of females (58,30 %). Although this might have been expected to favor model performance in the male group if sex had a significant influence, no such effect was observed in this study. A more likely factor affecting the performance between the sex-based subgroups is the slightly elevated AHI values of the male group compared to the female group, which is reflected in the slightly higher median, lower and upper quartile AHI values for the male subgroup. This theory is supported by the observed decrease in model performance with the increase of OSA severity.

For the OSA subgroups, the performance of both models changed with increasing OSA severity. Generally, the overall accuracies and Cohen's kappa values declined with the increase in OSA severity for both models. A similar trend was observed in a previous study, where the classification results were compared across OSA groups using a single EEG-channel (F4-M1) input for the automated analysis [10]. When examining the class-wise performance, a decline in the proportion of correctly detected REM sleep and N3 sleep was observed with the increase in OSA severity for the PG model analysis. While the PG model showed a more consistent decrease in its performance across OSA subgroups, a pronounced difference in the PSG model performance was observed for the severe OSA group compared to the other OSA severity-based groups. In this group, the PSG model demonstrated higher sensitivity for wake, indicating more of the wake stages were correctly identified, but also showed lower precision for wake due to misclassifying a greater proportion of sleep stages as wake. One likely explanation for the changes seen when shifting towards the more severe OSA categories is the increased number of breathing cessations during the night, which leads to greater sleep fragmentation [3]. This may, in turn, make the classification more challenging for the model. Additionally, the number of subjects included in the OSA severity-based subgroups decreased with increasing OSA severity. With a lower number of subjects individual recordings influence performance metrics more heavily.

The class-wise performance was very poor for stage N1 across all analyzed groups with only 4–5 % of N1 stages correctly classified. N1 was most frequently misclassified as N2 sleep and to a lesser extent, also as wakefulness, displaying a common misclassification pattern for N1 [19], [20]. The low agreement for scoring N1 was not a great surprise, as it has been established as the most difficult stage to classify for both humans [19], [20] and consequently automated analyses [5].

Both models showed a tendency to confuse other sleep stages with N2. N2 sleep typically constitutes the largest portion of total sleep, which was also the case in both the training dataset and the clinical dataset used in this study. Imbalances in the training dataset may cause the model to exhibit skewed performance towards the majority class if it is not very distinct from each other [5], in this case, N2 sleep. This tendency contributed to the decrease in specificity of detecting sleep, as also wake was relatively often misclassified as N2 sleep. However, as most of the analyzed epochs in the clinical dataset correspond to sleep, the specificity estimates are based on a smaller number of wake samples, which may impact the accuracy of the estimations. On the other hand, sensitivity for detecting sleep across all categories was primarily affected by N1 and N2 sleep being misclassified as wake. For the PG model analysis and the PSG model analysis of the severe OSA group, REM sleep was also relatively frequently confused as wake. Although only a relatively small proportion of N2 sleep was misclassified as wake (approximately 4 % in the PSG model analysis and 7 % in the PG model analysis), the high prevalence of N2 stages in the clinical dataset results in a substantial contribution to performance metrics. However, as only a modest portion of all sleep epochs were confused as wake, sensitivities of over 90 % for identifying sleep were reached across nearly all analyzed subgroups, with lower values reached only in the moderate and severe OSA groups. Overall, the sensitivity for detecting sleep was consistently higher than specificity across all analyzed categories, with the only exception being the PSG model analysis in the severe OSA group, suggesting that both models are generally more sensitive to sleep. For the sex-based groups, slightly higher sensitivities and specificities were observed in females compared to males for both models, further reflecting the slightly better performance of the deep learning models for females of this dataset. Among the OSA severity-based groups, the sensitivity decreased, while specificity increased progressively with decreasing OSA severity, highlighting a shift in the models' predictions towards wake.

Several factors could explain why the model did not perform overall as robustly on the clinical dataset as it could have been expected to based on the classification performance in the test data segmented from the training dataset. If the training dataset is not

representative of the specific population group the model is being applied to, a drop in performance is likely as the model encounters patterns it has not adequately learned during training [5]. The training dataset consisted of individuals with no suspected sleep abnormalities as well as those with suspected OSA. In contrast, the clinical dataset comprised individuals with suspected sleep-related abnormalities, including but not limited to OSA. It is possible that the training data was not representative enough to allow for the models to capture some of the sleep patterns it encountered in the clinical dataset. Additionally, although AASM defines scoring rules for sleep stages, the complexity of the sleep staging process and individual variation in the signals measured during sleep along with the inherent subjectivity introduce variation in the sleep staging results between scorers [20]. If the scoring patterns of the scorers in the training data vary between each other, the patterns available for the model to learn may be conflicting. It may also quite be possible that the single scorer of the clinical dataset has scoring patterns that differ from those learned by the model during training, which are in essence, scoring patterns influenced by the training dataset's scorers. It has been shown that the scorers from the same sleep center demonstrate a higher scoring agreement with each other compared to scorers from other sleep centers [19]. Considering just these factors alone, it was not altogether a great surprise that the models did not perform identically in the test dataset and the clinical dataset. Additionally, the model architecture itself can affect the model performance and generalizability [30]. For example, model architecture used in this study does not include a RNN network, which would allow to account for the information from previous time-steps [31] required in sleep staging [17]. Although Huttunen et al. [11], Rusanen et al. [7] and Tashakori et al. [9] have reported promising results for sleep staging using the same model architecture as was used in this study, it can be questioned whether this type of model is the optimal choice for sleep staging across different datasets.

As the aim was to examine the how the model would perform on the clinical dataset as-is, the clinical recordings were not specifically screened for artefacts or poor-quality signals. However, movement during recording, detached measuring electrodes, or a removed pulse oximeter can significantly impact the signal quality and consequently, the analysis results. The parts of the signal affected may, at worst, lose all relevant information required for scoring. Generally, several channels for sleep staging are available for the manual scoring and the human scorer can detect and discard any affected signals from their evaluation. As the models used in this study received only a limited number of signals and the models have limited ability to adapt to inputs they have not been trained

to recognize, their predictions can be vulnerable to disturbances in even one input channel. Using more signals as model inputs, pruning the affected segments from analysis or in some other way removing artifacts of the signals before model analysis seem like obvious approaches to mitigate these issues and attempt to improve the performance. However, care should be taken while choosing suitable methods for data pre-processing so that this phase does not undermine the time- and labor-saving benefits that automated analysis protocols aim to achieve and not to remove any information relevant to the classification. For example, as pointed out by Malafeev et al. [6], movement artifacts are commonly present in wakefulness and are also often related to transitions into stage N1, and may therefore be of relevance in the automated classification process.

It was also observed that the signals were affected by the Noxturnal version they were exported from and additionally, which recording device (A1 or A1S) they were recorded with. The exports from Noxturnal 6.3.0. compared to Noxturnal 7.0.0. had, for example, different quantized steps for some of the signals and different sampling rates for the same recordings. In addition, it seems that during the signal export, Noxturnal automatically filtered some signals with improper filters resulting in periodic artefacts that were especially visible in the SpO2-signals. These effects varied depending on the measurement device. In this study, the best compromise was determined to be to export all recordings using the latest available version (7.0.0) of Noxturnal. As the output of the model varies with the input, some variation in the results would likely have occurred with a different approach.

During deployment, a practical observation was made regarding computing environments that affected the model's predictions. Although the received models were deterministic and should theoretically produce consistent analysis results, variations were still observed between different machines and Python environments when testing the operation of the tools with test data before the analysis of the clinical data. This was pronounced when a custom, but optional, library was missing from the Python environment, leading to changes in the class-wise probabilities large enough to change some of the assigned sleep stages. Smaller shifts in the model predictions were observed between tests conducted on a Windows-based computer and those run in the Linux-based computing environment. We assume that this inconsistency was likely caused by floating-point errors, and potential minor variations in Python libraries between the two operating systems.

Comparing automated analysis results against a clinical evaluation may have some limitations, as the nature of clinical scoring can in practice vary from scorings conducted for research purposes. The purpose of clinical scoring is to confirm a diagnosis and evaluate

the severity of the disease to determine the prognosis and treatment options. Although accurate scoring of both sleep stages and respiratory-related events is essential for assessment, extremely precise scoring of “clear cases” (e.g., when severe OSA is evident) may only increase the labor intensity without providing significant additional benefit for clinical evaluation. This can reduce the agreement of the analysis between the clinician and the model. However, perhaps this effect can be expected to be more pronounced in the evaluation of respiratory-related events rather than in sleep staging, as sleep staging is conducted continuously over the entire analysis duration, whereas respiratory events are assessed as distinct occurrences throughout the recording.

The advantages of applying automated sleep staging are clear. The automated analysis, which can be conducted in seconds, removes the need for a manual repetitive and time-consuming task. Automated scoring methods could provide consistency and objectivity to sleep staging, while simultaneously reducing the time spent on the assessment and increasing the scalability of the analysis process without a proportional increase in resources. Additionally, they could make sleep scoring more accessible in settings where expert human scorers are not available. Moreover, automated sleep staging approaches which rely on unconventional signals such as PPG could improve the diagnostic value of measurements which lack the conventional means for sleep assessment, such as PG measurements. As PPG is a non-invasive measurement commonly included in health monitoring wearables, it could also enable cost effective long-term sleep monitoring. However, even despite the existence of commercially available automated analysis tools for sleep staging, they have not been widely adopted in clinical practice, with visual analysis remaining as the main approach in the clinical analysis of sleep [1], [5]. Several factors have been identified as possible contributors to the limited acceptance of the automated sleep staging tools [5], [8], some of which are highlighted also in the results of this study. The benefit of a new technology incorporated into the diagnostic process must provide value that is considered to exceed the perceived cost of incorporating it into the diagnostic process. As seen in this study, the models’ predictions in a dataset can differ from the analysis of a single scorer quite substantially and as discussed earlier in this section, this may be attributed to several different factors. Nevertheless, regardless of the underlying causes of the observed differences, an expert human scorer reviewing a model's output that substantially diverges from their own analysis would likely conclude that the model does not produce results that are reliable or practically useful for them. If the clinician must verify all sleep staging results manually, the expected advantages of the automated analysis process are effectively diminished. Moreover, as the clinician is ultimately responsible for the clinical evaluation, it may be challenging to rely

on deep learning technology especially since its decision-making process is not easily interpretable. Ultimately, deep learning models must be perceived as reliable not only by clinicians but also by patients to ensure trust and acceptance in clinical practice.

However, the main question of this study is not how the clinical acceptance of deep learning methods can be improved, but rather, are the methods currently good enough that they should be accepted for clinical use and in what capacity. It is reasonable to expect that a model's performance should reflect the variability observed among human scorers, particularly in deep learning approaches that rely on conventional sleep staging channels and are used to support clinical diagnosis. In such cases, discrepancies, especially in transitional or ambiguous sleep states, may be acceptable, as they mirror the known limitations of inter-scorer reliability rather than indicating model failure. However, although some disagreement in the sleep staging of the model and clinician is to be acceptable, this study shows that the deep learning models may not generalize optimally from dataset to dataset and may not perform consistently in patients with sleep disorders. A common limitation in existing literature is the reliance on test datasets derived from the training data, which restricts the ability to assess real-world performance. Therefore, further research is needed to evaluate model behavior across independent datasets and to identify the model architectures, training data, and data preprocessing and training strategies that can support robust generalization prior to clinical implementation before incorporation into clinical use.

One way to explore the possibilities of enhancing the model generalizability on specific datasets is through transfer learning, where a small subset of local data is used to fine-tune the model, allowing it to better adapt to the signal characteristics, patient population, and scoring practices of the sleep center in which it is applied. Another approach, which could perhaps better support the continuous day-to-day nature of clinical scoring, are human-in-the-loop approaches such as proposed by Rusanen et al. [7], wherein the model automatically flags uncertain or low-confidence epochs for manual review and fine-tunes the model with these manually scored segments using an active learning scheme. While the scoring agreement may drop especially in transitional states, regions of perfect agreement between several scorers also exist [19]. It has been shown that deep learning models tend to struggle with the same sleep epochs on which human scorers often disagree [7]. Thus, a human-in-the-loop approach could improve efficiency compared to sole manual in PSG scoring while simultaneously allowing for the clinicians to concretely leave their "fingerprint" in the analysis process, and to enhance the further analyses with the tweaked model.

The scope of this study was limited to performance examination at a group level, and as such it did not include any detailed individual-level examination. Many of the deep learning studies on sleep analysis also mainly evaluate the model performance on a group-level basis, with the results often tested in a test dataset segmented from the training dataset before training. However, while processing the analysis results of this study, variation in the model performance for individuals was observed, which is not captured in a group-level analysis. Further understanding of areas affecting the model performance in individuals is crucial to assess the suitability of deep learning tools in the clinical settings and across sleep disorders, and to determine areas for future improvement of these approaches.

In contexts where manual verification of automated analysis is not possible (i.e., conventional channels for sleep assessment are not available), such as in PG measurements, deep learning tools may serve a valuable role by enabling the screening of potential comorbid sleep disorders beyond the primary diagnostic focus such as OSA. Additionally, in certain clinical contexts, it may not be necessary to obtain detailed sleep stage information; rather, a general assessment of total sleep time and overall night-to-night sleep quality may be sufficient - particularly when compared to tools like questionnaires or actigraphy, which provide estimates that rely on subjectivity and may thus lack reliability. For these applications with no or only limited sleep-related information, deep learning tools even with lower accuracy may offer valuable insights without compromising the actual diagnosis. Furthermore, in conditions such as OSA, where the primary diagnostic metric, AHI, is influenced by total sleep time, simple accurate discrimination between sleep and wake could facilitate faster diagnosis. This may also help correct the known tendency of the REI, derived from PG, to underestimate OSA severity, as highlighted by Pitkänen et al. [13]. Models that are appropriately sensitive to sleep periods may better estimate total sleep time and, in turn, allow for a more representative calculation of the REI, aligning it more closely with the true AHI. However, this approach is not without limitations. Respiratory events occurring during wakefulness are typically excluded from analysis [17], which could become problematic if models increasingly classify segments as wake in patients with more severe OSA. Therefore, future research is needed to assess whether enhancing sleep–wake classification truly outperforms alternative strategies, such as adjusting REI thresholds, in improving diagnostic accuracy for OSA.

To conclude, the application of deep learning in clinical sleep analysis is still maturing toward clinical application. As Phan and Mikkelsen [8] have stated, challenges can be seen as opportunities for innovation. Each limitation encountered represents a stepping-stone toward more robust, adaptive, and clinically meaningful tools for sleep medicine.

6. CONCLUSIONS

This work evaluated the sleep staging performance of two deep learning models in an independent clinical dataset. While these methods hold significant promise for enhancing sleep analysis in clinical settings, the results of this study highlight that deep learning models do not always perform at the same level across datasets, highlighting the need for careful development and validation before clinical application. Future research should focus on identifying suitable model architectures, data preparation strategies, and training schemes that support robust performance across diverse datasets. Additionally, performance at the individual subject level should be more thoroughly assessed to better understand model behavior in real-world clinical scenarios. Ensuring that these models generalize well across varied patient populations and clinical environments is essential for their safe and effective use.

REFERENCES

[1] E. Mervaala, E. Haaksiluoto, S.-L. Himanen, S. Jääskeläinen, M. Kallio, and S. Vanhatalo, *Kliininen Neurofysiologia*, 1st ed. Helsinki: Duodecim, 2018.

[2] C. V. Senaratna *et al.*, 'Prevalence of obstructive sleep apnea in the general population: A systematic review', *Sleep Med. Rev.*, vol. 34, pp. 70–81, Aug. 2017, doi: 10.1016/j.smrv.2016.07.002.

[3] S. P. Patil, H. Schneider, A. R. Schwartz, and P. L. Smith, 'Adult Obstructive Sleep Apnea', *Chest*, vol. 132, no. 1, pp. 325–337, Jul. 2007, doi: 10.1378/chest.07-0040.

[4] D. J. Gottlieb and N. M. Punjabi, 'Diagnosis and Management of Obstructive Sleep Apnea: A Review', *JAMA*, vol. 323, no. 14, p. 1389, Apr. 2020, doi: 10.1001/jama.2020.3514.

[5] L. Fiorillo *et al.*, 'Automated sleep scoring: A review of the latest approaches', *Sleep Med. Rev.*, vol. 48, p. 101204, Dec. 2019, doi: 10.1016/j.smrv.2019.07.007.

[6] A. Malafeev *et al.*, 'Automatic Human Sleep Stage Scoring Using Deep Neural Networks', *Front. Neurosci.*, vol. 12, p. 781, Nov. 2018, doi: 10.3389/fnins.2018.00781.

[7] M. Rusanen *et al.*, 'Retrospective validation of automatic sleep analysis with grey areas model for human-in-the-loop scoring approach', *J. Sleep Res.*, p. e14362, Oct. 2024, doi: 10.1111/jsr.14362.

[8] H. Phan and K. Mikkelsen, 'Automatic sleep staging of EEG signals: recent development, challenges, and future directions', *Physiol. Meas.*, vol. 43, no. 4, p. 04TR01, Apr. 2022, doi: 10.1088/1361-6579/ac6049.

[9] M. Tashakori, M. Rusanen, T. Karhu, R. Huttunen, T. Leppänen, and S. Nikkonen, 'Optimal Electroencephalogram and Electrooculogram Signal Combination for Deep Learning-based Sleep Staging', *IEEE J. Biomed. Health Inform.*, pp. 1–8, 2025, doi: 10.1109/JBHI.2025.3541453.

[10] H. Korkalainen *et al.*, 'Accurate Deep Learning-Based Sleep Staging in a Clinical Population with Suspected Obstructive Sleep Apnea', *IEEE J. Biomed. Health Inform.*, pp. 1–1, 2019, doi: 10.1109/JBHI.2019.2951346.

[11] R. Huttunen *et al.*, 'A Comparison of Signal Combinations for Deep Learning-Based Simultaneous Sleep Staging and Respiratory Event Detection', *IEEE Trans. Biomed. Eng.*, vol. 70, no. 5, pp. 1704–1714, May 2023, doi: 10.1109/TBME.2022.3225268.

[12] S. Nikkonen *et al.*, 'Automatic Respiratory Event Scoring in Obstructive Sleep Apnea Using a Long Short-Term Memory Neural Network', *IEEE J. Biomed. Health Inform.*, vol. 25, no. 8, pp. 2917–2927, Aug. 2021, doi: 10.1109/JBHI.2021.3064694.

[13] M. Pitkänen *et al.*, 'Respiratory event index underestimates severity of sleep apnea compared to apnea-hypopnea index', *Sleep Adv.*, vol. 5, no. 1, p. zpad054, Jan. 2024, doi: 10.1093/sleepadvances/zpad054.

[14] H. Korkalainen *et al.*, 'Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea', *Sleep*, vol. 43, no. 11, p. zsaa098, Nov. 2020, doi: 10.1093/sleep/zsaa098.

[15] K. Kazemi *et al.*, 'Improved sleep stage predictions by deep learning of photoplethysmogram and respiration patterns', *Comput. Biol. Med.*, vol. 179, p. 108679, Sep. 2024, doi: 10.1016/j.compbiomed.2024.108679.

[16] R. Huttunen *et al.*, 'Assessment of obstructive sleep apnea-related sleep fragmentation utilizing deep learning-based sleep staging from photoplethysmography', *SLEEP*, vol. 44, no. 10, p. zsab142, Oct. 2021, doi: 10.1093/sleep/zsab142.

[17] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. L. Marcus, and B. V. Vaughn, *AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*, Version 2.0. Darien IL: American Academy of Sleep Medicine, 2012.

[18] M. H. Silber, L. E. Krahn, and T. I. Morgenthaler, *Sleep Medicine in Clinical Practice*, 0 ed. CRC Press, 2016. doi: 10.3109/9781616310059.

[19] S. Nikkonen *et al.*, 'Multicentre sleep-stage scoring agreement in the Sleep Revolution project', *J. Sleep Res.*, vol. 33, no. 1, p. e13956, Feb. 2024, doi: 10.1111/jsr.13956.

[20] R. S. Rosenberg and S. Van Hout, 'The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring', *J. Clin. Sleep Med.*, vol. 09, no. 01, pp. 81–87, Jan. 2013, doi: 10.5664/jcsm.2350.

[21] H. Danker-Hopfe *et al.*, ‘Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard’, *J. Sleep Res.*, vol. 18, no. 1, pp. 74–84, Mar. 2009, doi: 10.1111/j.1365-2869.2008.00700.x.
[22] M. Younes, J. Raneri, and P. Hanly, ‘Staging Sleep in Polysomnograms: Analysis of Inter-Scorer Variability’, *J. Clin. Sleep Med.*, vol. 12, no. 06, pp. 885–894, Jun. 2016, doi: 10.5664/jcsm.5894.
[23] R. G. Norman, I. Pal, C. Stewart, J. A. Walsleben, and D. M. Rapoport, ‘Interobserver Agreement Among Sleep Scorers From Different Centers in a Large Dataset’, *Sleep*, vol. 23, no. 7, pp. 1–8, Oct. 2000, doi: 10.1093/sleep/23.7.1e.
[24] J. P. Bakker *et al.*, ‘Scoring sleep with artificial intelligence enables quantification of sleep stage ambiguity: hypnodensity based on multiple expert scorers and auto-scoring’, *Sleep*, vol. 46, no. 2, p. zsac154, Feb. 2023, doi: 10.1093/sleep/zsac154.
[25] V. K. Kapur *et al.*, ‘Clinical Practice Guideline for Diagnostic Testing for Adult Obstructive Sleep Apnea: An American Academy of Sleep Medicine Clinical Practice Guideline’, *J. Clin. Sleep Med.*, vol. 13, no. 03, pp. 479–504, Mar. 2017, doi: 10.5664/jcsm.6506.
[26] M. J. Sateia, ‘International Classification of Sleep Disorders-Third Edition’, *Chest*, vol. 146, no. 5, pp. 1387–1394, Nov. 2014, doi: 10.1378/chest.14-0970.
[27] J. Park, H. S. Seok, S.-S. Kim, and H. Shin, ‘Photoplethysmogram Analysis and Applications: An Integrative Review’, *Front. Physiol.*, vol. 12, p. 808451, Mar. 2022, doi: 10.3389/fphys.2021.808451.
[28] T. Leppänen *et al.*, ‘Pulse Oximetry: The Working Principle, Signal Formation, and Applications’, in *Advances in the Diagnosis and Treatment of Sleep Apnea*, vol. 1384, T. Penzel and R. Hornero, Eds., in Advances in Experimental Medicine and Biology, vol. 1384. , Cham: Springer International Publishing, 2022, pp. 205–218. doi: 10.1007/978-3-031-06413-5_12.
[29] L. Alzubaidi *et al.*, ‘Review of deep learning: concepts, CNN architectures, challenges, applications, future directions’, *J. Big Data*, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.
[30] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org
[31] Y. LeCun, Y. Bengio, and G. Hinton, ‘Deep learning’, *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
[32] D. Ballabio, F. Grisoni, and R. Todeschini, ‘Multivariate comparison of classification performance measures’, *Chemom. Intell. Lab. Syst.*, vol. 174, pp. 33–44, Mar. 2018, doi: 10.1016/j.chemolab.2017.12.004.
[33] M. L. McHugh, ‘Interrater reliability: the kappa statistic’, *Biochem. Medica*, pp. 276–282, 2012, doi: 10.11613/BM.2012.031.
[34] S. Nikkonen, H. Korkalainen, J. Töyräs, and T. Leppänen, ‘STAR sleep recording export software for automatic export and anonymization of sleep studies’, *Sci. Rep.*, vol. 12, no. 1, p. 15859, Sep. 2022, doi: 10.1038/s41598-022-19892-0.
[35] R. Huttunen, M. Rusanen, S. Nikkonen, H. Korkalainen, and S. Kainulainen, ‘A harmonized and interoperable format for storing and processing polysomnography data’, 2024, *arXiv*. doi: 10.48550/ARXIV.2402.06702.
[36] Sosiaali- ja terveysministeriö, *Laki sosiaali- ja terveystietojen toissijaisesta käytöstä*. 2019.
[37] M. Perslev, M. H. Jensen, S. Darkner, P. J. Jennum, and C. Igel, ‘U-Time: A Fully Convolutional Network for Time Series Segmentation Applied to Sleep Staging’, 2019, *arXiv*. doi: 10.48550/ARXIV.1910.11162.
[38] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, ‘DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs’, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.
[39] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, ‘Squeeze-and-Excitation Networks’, 2017, *arXiv*. doi: 10.48550/ARXIV.1709.01507.
[40] J. R. Landis and G. G. Koch, ‘The Measurement of Observer Agreement for Categorical Data’, *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.