

PUBLICATIONS OF
THE UNIVERSITY OF EASTERN FINLAND



UNIVERSITY OF
EASTERN FINLAND

Dissertations in Forestry and Natural Sciences

HENRI KORKALAINEN

Deep Learning for Next-Generation Sleep Diagnostics

Sophisticated computational methods for more efficient and accurate assessment of sleep and obstructive sleep apnea

PUBLICATIONS OF THE UNIVERSITY OF EASTERN FINLAND
DISSERTATIONS IN FORESTRY AND NATURAL SCIENCES

N:o 386

Henri Korkalainen

DEEP LEARNING FOR NEXT-GENERATION SLEEP DIAGNOSTICS:

**SOPHISTICATED COMPUTATIONAL METHODS FOR MORE
EFFICIENT AND ACCURATE ASSESSMENT OF SLEEP AND
OBSTRUCTIVE SLEEP APNEA**

ACADEMIC DISSERTATION

To be presented by the permission of the Faculty of Science and Forestry for public examination in the Auditorium SN200 in Snellmania Building at the University of Eastern Finland, Kuopio on October 2nd, 2020, at 12 o'clock noon.

University of Eastern Finland
Department of Applied Physics
Kuopio 2020

Grano Oy
Kuopio, 2020
Editors: Pertti Pasanen, Jukka Tuomela,
Martti Tedre, Raine Kortet

Distribution:
University of Eastern Finland Library / Sales of publications
julkaisumyynti@uef.fi
<http://www.uef.fi/kirjasto>

ISBN: 978-952-61-3468-0 (print)
ISSNL: 1798-5668
ISSN: 1798-5668
SBN: 978-952-61-3469-7 (pdf)
ISSNL: 1798-5668
ISSN: 1798-5676 (pdf)

Author's address:

University of Eastern Finland
Department of Applied Physics
P.O. Box 1627, 70211
KUOPIO, FINLAND
Kuopio University Hospital
Diagnostic Imaging Center
KUOPIO, FINLAND
email: henri.korkalainen@uef.fi

Supervisors:

Adjunct Professor Timo Leppänen
University of Eastern Finland
Department of Applied Physics
KUOPIO, FINLAND
Kuopio University Hospital
Diagnostic Imaging Center
KUOPIO, FINLAND
email: timo.leppanen@uef.fi

Professor Juha Töyräs

The University of Queensland
School of Information Technology and
Electrical Engineering
BRISBANE, AUSTRALIA
University of Eastern Finland
Department of Applied Physics
KUOPIO, FINLAND
Kuopio University Hospital
Diagnostic Imaging Center
KUOPIO, FINLAND
email: juha.toyras@uef.fi

Adjunct Professor Sami Myllymaa

University of Eastern Finland
Department of Applied Physics
KUOPIO, FINLAND
Kuopio University Hospital
Diagnostic Imaging Center
KUOPIO, FINLAND
email: sami.myllymaa@uef.fi

Academic Fellow Isaac Afara

University of Eastern Finland
Department of Applied Physics
KUOPIO, FINLAND
The University of Queensland
School of Information Technology and
Electrical Engineering
BRISBANE, AUSTRALIA
email: isaac.afara@uef.fi

Reviewers:

Professor Maarten De Vos

University of Oxford
Institute of Biomedical Engineering
OXFORD, UNITED KINGDOM
KU Leuven
Department of Electrical Engineering
LEUVEN, BELGIUM
email: maarten.devos@eng.ox.ac.uk

Professor Sebastiaan Overeem

Eindhoven University of Technology
Department of Electrical Engineering
EINDHOVEN, THE NETHERLANDS
Kempenhaeghe
Sleep Medicine Center
HEEZE, THE NETHERLANDS
email: s.overeem@tue.nl

Opponent:

Gonzalo C. Gutiérrez-Tobal, Ph.D.

University of Valladolid
Department of Theory of Signal and Communications
and Telematic Engineering
VALLADOLID, SPAIN
email: gonzalo.gutierrez@gib.tel.uva.es

Henri Korkalainen

Deep Learning for Next-Generation Sleep Diagnostics: sophisticated computational methods for more efficient and accurate assessment of sleep and obstructive sleep apnea

Kuopio: University of Eastern Finland, 2020; 386

Publications of the University of Eastern Finland

Dissertations in Forestry and Natural Sciences

ABSTRACT

Currently, the diagnosis of sleep disorders relies on polysomnographic recordings with a time-consuming manual analysis with low reliability between different manual scorers. Throughout the night, sleep stages are identified manually in non-overlapping 30-second epochs starting from the onset of the recording based on electroencephalography (EEG), electrooculography (EOG), and chin electromyography (EMG) signals which require meticulous placement of electrodes. Moreover, the diagnosis of many sleep disorders relies on outdated guidelines. When assessing the severity of obstructive sleep apnea (OSA), the patients are classified based on thresholds of the apnea-hypopnea index (AHI), i.e. the number of respiratory disruptions during sleep. These thresholds are not fully based on solid scientific evidence and remain the same across different measurement techniques. The AHI does not correlate well with daytime symptoms and severe health outcomes. Moreover, OSA often leads to sleep fragmentation but its extent is often neglected in the diagnosis of OSA.

This thesis aimed to improve the diagnosis of sleep disorders by employing state-of-the-art computational and machine learning methods. The first aim was to simulate various AHI thresholds and optimize the severity classification with regards to OSA-related all-cause mortality. The second aim was to develop a comprehensive deep learning-based method for automatic sleep staging from a combination of EEG and EOG recordings, from a single-channel EEG, and finally, from a photoplethysmogram (PPG) measured with a finger pulse oximeter. The final aim was to implement the developed deep learning-based sleep staging to evaluate the sleep architecture in more detail to better identify sleep stage transitions automatically.

This thesis revealed that the current OSA severity classification is not optimal for assessing the risk for OSA-related all-cause mortality. Instead of the currently used AHI thresholds (5-15-30 h⁻¹) for mild, moderate, and severe OSA, the combination of 3-9-24 h⁻¹ would better reflect the risk of all-cause mortality when the AHI is determined from home-based polygraphic recordings. However, more detailed measures are required alongside the AHI for a comprehensive assessment of OSA severity. In the future, automated assessment of sleep fragmentation related to OSA and other respiratory event-based or hypoxemia-based parameters could supplement the severity estimation of OSA.

The developed deep learning-based sleep staging method was highly accurate with both the EEG+EOG combination and with a single frontal EEG channel. The methods achieved similar reliability as manual scoring in a clinical dataset of patients with suspected OSA. Moreover, deep learning enabled sleep staging with a moderate epoch-to-epoch agreement to manual sleep staging from a PPG

signal measured with a finger pulse oximeter and achieved a reasonably accurate determination of total sleep time. Deep learning further enabled a more detailed assessment of sleep architecture and sleep continuity. The more detailed approach enabled the deep learning-based sleep staging to better reveal the highly fragmented sleep architecture of individuals suffering from severe OSA.

In conclusion, the results of this thesis demonstrated that the severity assessment of OSA should be revised, sleep staging can be conducted fully automatically from even a single EEG channel or a photoplethysmogram and deep learning-based sleep staging may represent the solution for a more comprehensive assessment of sleep architecture. The results could significantly enhance the current diagnostic practice by making the analysis of sleep recordings more efficient and comprehensive while enabling simpler measurement setups and increasing the clinical usability and diagnostic value of simple home-based measurements.

National Library of Medicine Classification: W 26.55.A7, WG 141.5.P7, WL 108, WL 150

Medical Subject Headings: Sleep; Sleep Stages; Dysomnias/diagnosis; Sleep Apnea, Obstructive/diagnosis; Machine Learning; Deep Learning; Polysomnography; Photoplethysmography; Oximetry; Electroencephalography; Electrooculography

Yleinen suomalainen asiasanasto: uni (lepotila); unitutkimus; unihäiriöt; uniapnea-oireyhtymä; tekoäly; koneoppiminen; EEG

ACKNOWLEDGEMENTS

This thesis was carried out at the Department of Applied Physics, University of Eastern Finland and Diagnostic Imaging Center, Kuopio University Hospital during the years 2018-2020 and was financially supported by the Research Committee of the Kuopio University Hospital Catchment Area for the State Research Funding (project numbers 5041780 and 5041767), the Respiratory Foundation of Kuopio Region, the Research Foundation of the Pulmonary Diseases, the Foundation of the Finnish Anti-Tuberculosis Association, the Päivikki and Sakari Sohlberg Foundation, the Veritas Foundation, and the Academy of Finland (grant number 313697). I would like to thank all the involved parties

Firstly, I'd like to thank the supervisors of this thesis: Adjunct Professor Timo Leppänen, Professor, Chief Physicist Juha Töyräs, Adjunct Professor Sami Myllymaa, and Academic Fellow Isaac Afara. Thank you for all your support and help. You have always given thorough and helpful comments and I consider myself privileged for having such dedicated and hard-working supervisors! I would also like to thank the external reviewers of this thesis, Professor Maarten De Vos and Professor Sebastiaan Overeem. Furthermore, I'd like to express my gratitude to Gonzalo Gutiérrez-Tobal for agreeing to act as my opponent.

I owe my warmest gratitude to all the co-authors. Firstly, I'd like to thank Juhani Aakko; discussing ideas and sharing codes with you has been extremely useful and this thesis would not be the same without you. Secondly, I offer my sincere gratitude to Brett Duce. This thesis would not have been possible without you. Thank you for providing all the necessary data for the studies and ensuring that it is of the highest quality. Discussing ideas with you has been extremely helpful and lastly, I'd like to thank you for always greeting me with a cup of coffee (the best coffee I had in Brisbane!).

I would also like to thank the remaining co-authors, everyone belonging to the Sleep Technology and Analytics Group, and my fellow researchers at the hospital for all the lunch and coffee breaks along with all the helpful discussions related (and not so related) to research. A special thanks goes to Samu and Sami; thank you both for sharing an office with me at some point and for all the helpful conversations, advice, and just listening to any worries. You have both been a huge help during this thesis. Finally, I'd also like to express my gratitude to Ewen MacDonald for thoroughly proofreading the thesis and to Tuomas Lunttila for all your help regarding the servers, computers, and remote use.

I want to extend my deepest gratitude to my parents. You have always supported me and I really couldn't hope for better parents. I'd also like to thank my friends and family for ensuring that not everything I do is related to work. Thank you for all the climbing sessions, going to the gym, playing tennis, playing Smash Bros or Mario Kart, and all the other numerous activities and helpful distractions. One more thank you is owed that has tremendously helped me during this project and life in general: thank you Matti and Heikki for introducing me to playing the guitar all those years ago. I'm not sure how I would have managed and kept sane without my guitars during the writing of this thesis as it always seemed to help me whenever I felt stuck and most of the best ideas always seem to come to me during playing.

Finally, I would like to extend my heartfelt and deepest gratitude to my amazing wife Minna. You have made sure that there's more in life than just work and you have made my life better in every aspect. Thank you for always being there for me, for all the good times we've had, and for all the good times that are ahead.

"One, remember to look up at the stars and not down at your feet. Two, never give up work. Work gives you meaning and purpose and life is empty without it. Three, if you are lucky enough to find love, remember it is rare and don't throw it away"

-Stephen Hawking

Kuopio, August 25, 2020

A handwritten signature in black ink, reading "Henri Korkalainen". The script is fluid and cursive, with the first name "Henri" written in a larger, more prominent hand than the last name "Korkalainen".

Henri Korkalainen

LIST OF ABBREVIATIONS

AASM	American Academy of Sleep Medicine
AHI	Apnea-hypopnea index
AMI	Acute myocardial infarction
API	Application programming interface
ANS	Autonomic nervous system
BMI	Body mass index
CBT	Cognitive behavioural therapy
CPAP	Continuous positive airway pressure
CSA	Central sleep apnea
CVD	Cardiovascular disease
CNN	Convolutional neural network
ECG	Electrocardiography
EEG	Electroencephalography
EMG	Electromyography
EOG	Electrooculography
HRV	Heart rate variability
HSAT	Home sleep apnea test
ICSD	International classification of sleep disorders
IQR	Interquartile range
GRU	Gated recurrent unit
LSTM	Long short-term memory network
MSLT	Multiple sleep latency test
N1	N1 sleep stage (light sleep)
N2	N2 sleep stage (light sleep)
N3	N3 sleep stage (deep sleep)
NREM	Non-rapid eye movement sleep
OSA	Obstructive sleep apnea
PSG	Polysomnography
PG	Polygraphy
PPG	Photoplethysmography
REI	Respiratory event index
ReLU	Rectified linear unit
REM	Rapid eye movement sleep
RNN	Recurrent neural network
SGD	Stochastic gradient descent
SE	Sleep efficiency
SD	Standard deviation
SpO ₂	Saturation of peripheral oxygen
TST	Total sleep time
TRT	Total recording time
WASO	Wake after sleep onset

Throughout this thesis, light sleep denotes the combination of N1 and N2 sleep while deep sleep denotes N3 sleep. NREM sleep denotes N1, N2, and N3 sleep.

LIST OF SYMBOLS

a	The activation of a single layer of neurons in a neural network
α	Learning rate
b	Biases of the neural network
$C(\cdot)$	Cost function
$C_{\text{MSE}}(\cdot)$	Mean squared error cost function
$C_{\text{CE}}(\cdot)$	Cross-entropy cost function
f_s	Sampling frequency
$h^{(t)}$	The state of the hidden units of a recurrent neural network at a time t
κ	Cohen's kappa coefficient
∇	Gradient
n	Number of samples/patients
p	Probability to reject the correct null hypothesis
\mathbb{R}	Set of real numbers
\mathbb{R}^n	Real coordinate space of dimension n
σ	A non-linear activation function
σ_S	The sigmoid function
θ	Parameter values optimized during the training of a neural network
U	Input weights of a recurrent neural network
w	Weights of the neural network
W	Recurrent weights of a recurrent neural network
x	Input to the neural network
y	Output of the neural network

LIST OF ORIGINAL PUBLICATIONS

This thesis comprises a review of the author's work in the field of sleep medicine and biomedical engineering and informatics. The following publications are referred to by the Roman numerals I-IV.

- I Korkalainen H., Töyräs J., Nikkonen S., and Leppänen T. Mortality-risk-based apnea-hypopnea index thresholds for diagnostics of obstructive sleep apnea, *Journal of Sleep Research*, 28(6): e12855, 2019. doi: 10.1111/jsr.12855
- II Korkalainen H., Aakko J., Nikkonen S., Kainulainen S., Leino A., Duce B., Afara I.O., Myllymaa S., Töyräs J., and Leppänen T. Accurate Deep Learning-Based Sleep Staging in a Clinical Population with Suspected Obstructive Sleep Apnea, *IEEE Journal of Biomedical and Health Informatics*, 24(7): 2073-2081, 2019. doi: 10.1109/JBHI.2019.2951346
- III Korkalainen H., Aakko J., Duce B., Kainulainen S., Leino A., Nikkonen S., Afara I.O., Myllymaa S., Töyräs J., and Leppänen T. Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea, *Sleep*, zsa098, 2020. doi:10.1093/sleep/zsa098
- IV Korkalainen H., Leppänen T., Duce B., Kainulainen S., Aakko J., Leino A., Kalevo L, Afara I.O., Myllymaa S., and Töyräs J. Detailed assessment of sleep architecture with deep learning reveals sleep fragmentation of patients with obstructive sleep apnea better than traditional scoring (under review).

AUTHOR'S CONTRIBUTION

The publications included in this thesis were made in a collaboration between the Department of Applied Physics, University of Eastern Finland and the Diagnostic Imaging Center, Kuopio University Hospital with contributions from the Sleep Disorders Centre, Princess Alexandra Hospital (Brisbane, Australia), and the School of Information Technology and Electrical Engineering, the University of Queensland (Brisbane, Australia).

The author contributed to studies **I-IV** as follows:

- I** The author designed the study with the supervisors, participated in the study conception, was responsible for the data analyses, interpreted the results with the co-authors, and was the main writer of the manuscript.
- II** The author designed the study with the supervisors, participated in the study conception, was responsible for data analyses and development of the deep learning approaches, interpreted the results with the co-authors, and was the main writer of the manuscript.
- III** The author devised the main conceptual ideas of the study with J. Aakko and carried out the analyses and the writing of the manuscript in co-operation with co-authors. Together with the supervisors, the author was responsible for the study design and conception.
- IV** The author designed the study with B. Duce and J. Töyräs, was responsible for the data-analyses, interpreted the results with the co-authors, and was the main writer of the manuscript.

In all manuscripts, the collaboration with the co-authors has been significant.

TABLE OF CONTENTS

1	Introduction	1
2	Sleep	3
2.1	Sleep architecture	3
2.2	Sleep disorders	5
2.2.1	Obstructive sleep apnea	5
2.2.2	Other sleep disorders	7
2.3	Biosignal recordings	9
2.3.1	Polysomnography	9
2.3.2	Portable sleep monitoring	11
2.3.3	Actigraphy	12
3	Deep learning	13
3.1	Fully connected feedforward neural networks	13
3.1.1	Convolutional neural networks.....	16
3.1.2	Recurrent neural networks.....	17
3.2	Applications in medicine.....	20
4	Aims of the thesis	23
5	Methods	25
5.1	Study populations and measurement devices.....	25
5.2	Optimizing the AHI thresholds used for OSA severity classification	28
5.3	Deep learning-based sleep staging.....	30
5.3.1	Neural network architecture	30
5.3.2	The training process and performance evaluation.....	32
5.4	Deep learning-based sleep staging with better temporal resolution	33
5.5	Statistical analyses.....	35
6	Results	37
6.1	Mortality risk-based AHI thresholds for OSA severity classification	37
6.2	Deep learning-based automatic sleep staging based on EEG and EOG recordings	42
6.2.1	Sleep staging in a public dataset of healthy individuals	42
6.2.2	Sleep staging in a clinical dataset of patients with suspected OSA.....	42
6.2.3	Effect of OSA severity on sleep staging	44
6.3	Deep learning-based automatic sleep staging based on photoplethysmogram.....	46
6.3.1	Sleep staging accuracy.....	46
6.3.2	Derived clinical parameters	49
6.4	Detailed analysis of sleep architecture with deep learning	51

6.4.1	Sleep stage percentages and sleep parameters	51
6.4.2	Assessing sleep fragmentation via survival analysis.....	54
7	Discussion	57
7.1	Optimizing the severity assessment of OSA.....	57
7.2	Deep learning-based sleep staging.....	59
7.3	Detailed analysis of sleep architecture.....	62
8	Conclusions	65
	BIBLIOGRAPHY	67

1 Introduction

Sleep is a restorative state essential for physical and mental recovery, memory consolidation, and the clearance of metabolic waste products from the brain [1,2]. However, sleep disorders fragment sleep and decrease sleep quality leading to excessive daytime sleepiness, impaired vigilance, and various severe health consequences. One of the most common sleep disorders is obstructive sleep apnea (OSA) characterized by recurrent obstructions of the upper airways during sleep [3,4]. These breathing disruptions often lead to recurrent hypoxemic periods and arousals from sleep evoking a significant decline in sleep quality and daytime vigilance [5,6]. Furthermore, OSA and poor sleep quality are related to decreased quality of life, an increased risk of traffic accidents, and various comorbidities, such as cardiovascular diseases [7–9]. Inadequate sleep and sleep disorders are a major global health problem affecting a large portion of the world’s population and posing a significant economical burden induced by the related comorbidities, accidents, and loss of productivity [10].

Despite the high prevalence of sleep disorders, the current diagnostic practice relies on a time-consuming and labour-intensive manual analysis of an overnight, in-laboratory recording, polysomnography (PSG). From the PSG, sleep stages are identified to assess the sleep architecture. Currently, the goal is to manually segment the night into 30-second epochs with a single sleep stage identified for each epoch. Sleep is categorized into rapid eye movement (REM) sleep and into three non-REM (NREM) stages, two of which are considered light sleep (stages N1 and N2) and a deep sleep stage (N3) [11]. However, as the current manual analysis is based on multiple recorded signals of electroencephalography (EEG), electrooculography (EOG), and chin electromyography (EMG), it can take several hours to analyse the signals from a single patient. Moreover, the arbitrary division of the night into 30-second epochs with only a single representative sleep stage for each epoch may cause several transitions between sleep stages being overlooked. This can be a serious problem when diagnosing sleep disorders. Moreover, the division was developed based on the sleep of healthy individuals and is a historical remnant from an era when each 30-second period of the recorded signals was printed on paper [12,13]. Nonetheless, these outdated practices form the cornerstone for clinical diagnosis of sleep disorders.

Especially in the diagnosis of OSA, home-based ambulatory polygraphies (PG) are often used instead of a PSG. The greatest limitation of a PG recording is the lack of EEG, EOG, and chin EMG [14]. Thus, a manual sleep staging is impossible and this prevents the assessment of sleep architecture and the diagnosis of other comorbid sleep disorders. Furthermore, counting the number of breathing disruptions forms the basis of the diagnosis and the severity assessment of OSA. Both the complete cessations in breathing (apneas) and partial obstructions (hypopneas) are combined into a single metric, the apnea-hypopnea index (AHI). Based on a set of AHI thresholds, the OSA severity is defined and this assessment often dictates which patient is eligible to receive subsidized treatment [3]. The sensitivity of the recordings setups has significantly improved over the years and

the definitions of hypopneas have varied enabling more events to be identified. Despite these developments, the thresholds used to assess the severity of OSA have remained unchanged [3, 15].

Deep learning has already been applied to assist in automatic detection and classification of medical conditions [16, 17]. Deep learning is a machine learning technique based on multiple layers of artificial neural networks designed to mimic the biological function of neurons. While the traditional programming paradigm relies on explicitly stating a solution to a problem with a set of rules, deep learning is based on automatically learning rules and patterns from a set of examples. This makes it possible to find solutions to highly complex problems. Deep learning has already revolutionized tasks such as speech recognition and image classification [18, 19].

The research included in this thesis was conducted to enhance and optimize the diagnosis of sleep disorders, with an emphasis on OSA, and provide methods for automatically and accurately identifying the sleep stages. The research aimed to optimize the severity classification of OSA so that it would better correspond to the risk of severe health consequences. Moreover, the aim was to develop deep learning-based approaches for sleep staging from lighter measurement setups than a full PSG with either a single EEG channel or a photoplethysmogram (PPG) measured with a finger pulse oximeter. Finally, we aimed to move beyond the current sleep staging practice restricted by the non-overlapping 30-second epochs by analysing sleep architecture with better temporal resolution. One hypothesis was that by optimizing the AHI thresholds used to assess OSA severity, this would achieve a better differentiation of patients with an elevated risk of OSA-related health consequences. Furthermore, we hypothesised that a deep learning approach would permit sleep staging with a single EEG channel and that sleep staging could be conducted by relying on PPG. Finally, we hypothesised that a more detailed analysis of sleep architecture with deep learning not restricted to non-overlapping 30-second epochs would provide a better assessment of OSA-related sleep fragmentation. We anticipated that with the approaches implemented in this thesis, we could achieve an optimized severity classification of OSA as well as readily implementable automatic methods for sleep staging. In this way, these methods could reduce the clinical workload and improve the diagnostic yield of ambulatory recordings. The final goal was to devise a novel automatic method capable of assessing sleep architecture with a better temporal resolution.

2 Sleep

Sleep can be considered as a reversible mental and physical state characterized by a lack of physical activity and a degree of unresponsiveness to the environmental stimuli. However, sleep is not simply the absence of wakefulness; rather, it has its own internal structure, so-called sleep architecture [11,12]. Sleep is not a constant state as there exists variation between different distinct stages following a typical temporal structure [20,21]. However, various sleep disorders can disturb the natural sleep architecture causing insufficient and non-restorative sleep further disrupting daily functioning and causing significant health consequences [5,22].

The underlying reason why we spend such a large portion of our lifetime asleep remains largely unknown even though many important functions of sleep have been discovered. Overall, sleep is a highly restorative state, both physically and mentally. Sleep is essential for memory consolidation [1], learning and strengthening of cognitive skills [23], and the recovery and growth of muscles [24]. Furthermore, sleep, and especially deep sleep, allows the brain to clear out excess metabolic waste [2]. Conversely, sleep deprivation causes adverse mental and physiological effects such as impairment of short- and long-term memory [25], alterations in immunological defence [26], and deterioration of cognitive performance [27]. Moreover, sleep deprivation and untreated sleep disorders have been linked to depression [28], cardiovascular disease [29], and mood disorders [30].

The following chapters present the basic concepts of sleep architecture and explain how sleep and sleep disorders can be assessed from biosignal recordings. The main focus of this thesis is the diagnosis of obstructive sleep apnea. However, the diagnostic recordings are universal to various sleep disorders. Therefore, a brief overview of different sleep disorders and their diagnostic approaches is presented to achieve a more comprehensive representation and to fully illustrate the potential of the methods developed in this thesis.

2.1 SLEEP ARCHITECTURE

Sleep can roughly be divided into three distinct periods: wakefulness, REM (rapid eye movement) sleep, and NREM (non-rapid eye movement) sleep. Furthermore, NREM sleep can further be divided into three stages: N1 and N2 sleep comprising light sleep, and N3 sleep considered as deep sleep. Previously, deep sleep was further divided into two distinct stages according to the classification of Rechtschaffen and Kales [12] but this practice has been abandoned in the current clinical practice based on the guidelines issued by the American Academy of Sleep Medicine (AASM) [11].

The sleep architecture is assessed via sleep staging. This involves identifying the sleep stages from recordings of electroencephalography (EEG), electrooculography (EOG), and chin electromyography (EMG) assessing the electrical activity of the brain, movement of the eyes, and chin muscle tone, respectively. According to current practice, sleep stages are identified in consecutive 30-second epochs and a single stage is assigned for each epoch [11].

In the sleep staging, the frequency content of EEG is divided into five categories. Based on the EEG frequency f , the categories are: 1) slow-wave activity: $0.5 \leq f \leq 2.0$ Hz with $>75 \mu\text{V}$ peak-to-peak amplitude in the frontal EEG channels; 2) delta waves: $0 < f < 4.0$ Hz; 3) theta waves $4.0 \leq f < 8.0$ Hz; 4) alpha waves: $8.0 \leq f < 13.0$ Hz; and 5) beta waves $f \geq 13.0$ Hz [11].

Wakefulness is characterized by an alpha rhythm, i.e. trains of alpha waves, in the EEG when the eyes are closed. With eyes open, EEG activity comprises both alpha and beta waves with a low amplitude and without the same rhythmicity as with closed eyes. However, some individuals fail to generate an alpha rhythm or do so to a limited extent and thus no major differences can be detected between the EEG activity with eyes open or closed. Moreover, it is common during wakefulness for eye blinks with conjugate vertical eye movements to occur; these can be detected in the EOG at a frequency of around 0.5-2 Hz. Furthermore, even rapid eye movements may be present, but the muscular tone in the chin EMG remains high, differentiating these movements from those evident in REM sleep. Finally, slow eye movements may occur during wakefulness but also during N1 sleep [11].

In addition to slow eye movements, the first light sleep stage, N1 sleep, is characterised by low-amplitude, mixed frequency EEG activity predominately in the theta frequencies. For most individuals, N1 sleep is the first occurring sleep stage after wakefulness and defines the sleep onset. As for the chin muscle tone, N1 sleep still has varying chin EMG amplitudes. However, the amplitudes are generally lower than those encountered during wakefulness [11]. With the onset of N1 sleep, conscious awareness of the environment slowly decreases [31]. However, the arousal threshold remains relatively low during N1 sleep and thus external or internal stimuli can easily lead to awakening [20].

The second stage of light sleep, N2, can be differentiated from N1 sleep by the occurrence of K-complexes and sleep spindles which are characteristic to the N2 stage. K-complex is a sharp wave with both negative and positive components whereas a sleep spindle is a train of sinusoidal waves of 11 – 16 Hz frequency. Both the K-complex and sleep spindle are identified from the EEG and must have a duration of ≥ 0.5 seconds [11]. During N2, the EOG generally does not illustrate any eye movement activity; however, some individuals still retain the slow eye movements. As for the chin EMG, the amplitude varies and is usually lower than during wakefulness [11]. In contrast to N1 sleep, N2 is characterized by a complete disappearance of conscious awareness [31] and the arousal threshold is higher [20].

The deep sleep stage, N3, is characterized by slow-wave activity ($0.5 \leq f \leq 2.0$ with amplitude $>75 \mu\text{V}$) visible in the EEG [11]. There are generally no visible eye movements in the EOG during N3 sleep and thus the EOG signal usually only displays the same frequencies as the EEG. Moreover, the chin EMG amplitude may vary, but it is generally lower than during wakefulness and N2 sleep [11]. N3 is the deepest sleep stage with no conscious awareness and is the most difficult stage from which to be awakened [20, 31]. N3 is important for memory consolidation [1, 32] and is essential for the clearance of metabolic waste products from the brain via cerebrospinal fluid flow [2].

REM sleep is characterized by rapid eye movements (initial deflection < 500 ms in the EOG) resembling those when visually scanning the environment during wakefulness [11]. These are visible in both EOG channels as concurrent out-of-phase deflections. The EEG pattern during REM sleep is highly similar to wakefulness but can illustrate sawtooth waves which are trains of sharp, 2–6 Hz waves with high amplitude [5, 11]. During REM sleep, transient muscle activity may occur

and is visible as short bursts of chin EMG activity (<0.25 s); however, the muscle tone is the lowest of all sleep stages [11]. REM is important for learning and memory consolidation, especially of procedural and motoric skills and dreaming also commonly occurs during REM sleep [20,33–36].

The transition between sleep stages during normal sleep usually occurs in cycles. First, sleep gradually deepens from N1 and N2 to N3 before first transitioning to REM sleep after about 70 minutes from sleep onset [33]. After the REM period, the sleep cycle is repeated and the REM periods occur about 80–120 minutes after the end of the previous REM period [20, 33, 34]. The first REM period is usually the shortest with a typical duration of around 5 minutes but the durations increase during the night [20,33]. Conversely, the duration of continuous N3 periods decreases throughout the night [5,20]. Usually, the N3 sleep occurs during the first sleep cycles, most likely due to the high importance of N3 sleep. Moreover, the duration of N3 sleep increases after sleep deprivation [20]. Generally, N2 accounts for most of the sleep, typically around 45–50% of the total sleep time. N1 usually comprises less than 5% of sleep while N3 represents around 20–25% [20, 34]. REM usually is responsible for approximately 20–25% of sleep while approximately 5% of the time between sleep onset and awakening in the morning is spent awake [5,20,34].

Even though sleep stages are defined based on the frequency content in EEG, they are also reflected in the activity of the autonomic nervous system. When progressing from wakefulness to deep sleep, the parasympathetic tone increases progressively while the sympathetic tone decreases [37, 38]. Conversely, REM sleep typically is accompanied by an increased sympathetic tone and decreased parasympathetic tone [39]. The periods of wakefulness during the night have a sympathetic and parasympathetic tone between NREM and REM sleep [40].

2.2 SLEEP DISORDERS

Sleep disorders are divided into six main categories according to the *International Classification of Sleep Disorders (ICSD)*: sleep-related breathing disorders, insomnia disorders, circadian rhythm sleep-wake disorders, central disorders of hypersomnolence, parasomnias, and sleep-related movement disorders [41]. In the following section, the most common sleep-related breathing disorder, obstructive sleep apnea, is described. After this, a brief overlook is given on the remaining five sleep disorder categories.

2.2.1 Obstructive sleep apnea

Obstructive sleep apnea (OSA) is a highly prevalent sleep-related breathing disorder affecting up to 900 million individuals globally [4]. OSA is characterized by recurrent respiratory disruptions during the night. Partial obstructions of the upper airways are called hypopneas while complete cessations in breathing are called apneas [3]. OSA can cause a significant disruption to sleep quality due to the recurrent arousals from sleep caused by the respiratory disruptions [3]. Individuals suffering from OSA have, in general, a more fragmented sleep architecture and less deep sleep during the night [5, 42]. OSA is also related to various daytime symptoms; for example, excessive daytime sleepiness and impaired vigilance [6,43]. Furthermore, individuals suffering from OSA generally have a higher risk for traffic accidents, cardiovascular disease, cancer, stroke, and all-cause mortality [7–9,44–46]. OSA represents not only a major healthcare burden and significant economical

costs but also indirectly via the downstream health sequelae [10].

An OSA diagnosis mainly relies on an overnight polysomnography (PSG) [11]. However, most likely due to the limited availability and high cost of PSG, many individuals affected by OSA remain undiagnosed and without treatment [47]. It has been estimated that 80% of individuals affected with OSA remain untreated in the USA [48]. This can be a major issue as undiagnosed OSA significantly elevates the healthcare costs [49]. Moreover, the elevated costs can usually be reduced to the same level as the general population by successful treatment and the treatment could also significantly improve the quality of life for the affected individuals [50]. It has been estimated that undiagnosed OSA results in over \$6000 annual costs per person but can be reduced to around \$2000 after treatment [48]. To overcome the limitations related to the attainability of PSG, ambulatory polygraphies (PG) are occasionally used in the diagnosis of OSA and are even the preferred diagnostic method in some healthcare systems [51]. However, PG lacks recording of EEG, impeding the assessment of sleep architecture [14]. A more efficient and comprehensive diagnosis of OSA without having to rely on an in-lab PSG would be essential to alleviate the high healthcare burden.

Apnea is defined as an event where the airflow signal amplitude decreases by over 90% from the baseline and this lasts for ≥ 10 seconds. Conversely, hypopnea is defined as a $\geq 30\%$ decrease in the airflow signal amplitude for ≥ 10 seconds [11]. Furthermore, hypopnea must be associated with an arousal from sleep or a $\geq 3\%$ decrease in oxygen saturation [11]. However, there have been several definitions produced for identifying hypopneas over the years [3, 11, 15]. Previously, hypopnea had to be associated with a $\geq 4\%$ decrease in oxygen saturation [3] and this definition remains an acceptable alternative [11]. However, the desaturation threshold significantly affects the number of identified hypopneas and the 3% desaturation threshold has led to significantly more hypopneas being identified [52, 53]. It is recommended that apneas are identified using an oronasal thermal airflow sensor to detect the reduction in airflow whereas a nasal pressure transducer is used for detecting hypopneas [11].

The main diagnostic parameter to assess the severity of OSA and the necessity of treatment is the apnea-hypopnea index (AHI) [3, 11]. The AHI is calculated from the overnight recordings as the number of apneas and hypopneas normalized by the total sleep time or total recording time [3]. Total sleep time is used with PSG while the total recording time is used with PG as the determination of the total sleep time is impossible with the conventional manually conducted visual assessment of EEG. The term respiratory event index (REI) is also used to refer to the AHI derived from PG [11]. Moreover, arousals from sleep are not identified with the current PG analysis methods leading to the fact that all of the hypopneas associated only with an oxygen desaturation are counted while those linked with an arousal from sleep remain overlooked. Due to these reasons, the AHI values determined based on PSG and PG can differ significantly [52, 54]. However, in both PSG and PG, the OSA severity is classified based on the same thresholds of AHI: $5 \text{ h}^{-1} < \text{AHI} \leq 15 \text{ h}^{-1}$ indicates mild OSA, $15 \text{ h}^{-1} < \text{AHI} \leq 30 \text{ h}^{-1}$ indicates moderate OSA, while $\text{AHI} \geq 30 \text{ h}^{-1}$ indicates severe OSA [3]. Regardless of large differences between the AHI derived from PSG and PG, the same AHI thresholds are always used even though these lack strong scientific foundations and clinical evidence [54, 55].

The most commonly used treatment for OSA is continuous positive airway pressure therapy (CPAP) [56]. However, while CPAP is highly effective in preventing the respiratory events and can improve daytime functioning and decrease sleepiness,

the adherence is low, most likely due to its sleep-disrupting nature (e.g. noise, uncomfortable fitting, and sweating under the mask) [57]. Weight loss can also assist in managing OSA and reduce the number of respiratory events [56,58]. Furthermore, when the majority of the respiratory events occur in the supine position, positional therapy may be used to prohibit supine position [56,59]. Moreover, mandibular devices or surgical approaches are also used [56] and hypoglossal nerve stimulation has produced promising results [60]. A few pharmacological interventions also exist but are mainly focused on treating the excessive daytime sleepiness related to OSA [56].

2.2.2 Other sleep disorders

Aside from OSA, another common sleep-related breathing disorder is central sleep apnea (CSA). The main difference between OSA and CSA is the occurrence of central apneas. Central apneas are characterized by a lack of effort to begin breathing during the respiratory disruptions [61]. The diagnosis of central sleep apnea follows the same procedure as OSA, and CSA can be differentiated from OSA based on PSG or ambulatory PG. The treatment of CSA relies on supplemental oxygen or treating the associated medical problems that may contribute to CSA (e.g. treating heart failure or reducing opioid-based medications) [61,62]. Similarly to the situation with OSA, CPAP may also occasionally alleviate the symptoms [62].

Insomnia is characterized by difficulties in falling asleep, maintaining sleep, or early awakenings. Insomnia is related to poor sleep quality with a short total sleep time not explained by environmental factors and restrictions [22]. Short-term insomnia can occasionally occur in up to half of the adult population, while insomnia together with daytime impairment occurs in 10 to 15% of the population [63]. The diagnosis of insomnia is based on questionnaires assessing comorbid disorders and daytime dysfunction together with sleep logs, sleep diaries, and actigraphy recordings [63]. According to current practices, a PSG is only used when other sleep disorders, such as sleep apnea, are also suspected or when the diagnosis or treatment otherwise is inconclusive or insufficient [63,64]. The most common treatment for insomnia is cognitive behavioural therapy (CBT) but pharmacological interventions are used if CBT is not effective [64]. Insomnia often co-occurs with OSA [65]; however, comorbid insomnia often remains overlooked when OSA is diagnosed without a PSG-based analysis of sleep architecture.

Circadian rhythm sleep-wake disorders arise from misalignment of the sleep-wake cycle in relation to the environment and the light-dark cycle. These may be either caused by intrinsic factors (e.g. non-24h sleep-wake rhythm and advanced or delayed sleep-wake phase) or by extrinsic, environmentally induced misalignments (e.g. shift work and jet lag disorders) [41,66]. The main diagnostic method to assess circadian rhythm sleep-wake disorders is actigraphy and various biomarkers such as melatonin secretion onset in dim-light conditions [41,67]. Commonly, these disorders are treated with either strategically timed melatonin administration, light therapy, or behavioural interventions [66].

Central disorders of hypersomnolence are mainly caused by abnormalities in the central nervous system and in controlling the sleep-wake balance [41]. These manifest as excessive daytime sleepiness despite a normal timing and quality of sleep and cannot be related to being caused by another sleep disorder [41,68]. For example, type 1 and 2 narcolepsy and idiopathic hypersomnia are all characterized

as an irrepressible need to sleep and excessive daytime sleepiness [41, 68, 69]. The diagnosis of central disorders of hypersomnolence mainly relies on a multiple sleep latency test (MSLT) which assesses the extent of excessive daytime sleepiness. Moreover, actigraphy and sleep diaries are occasionally used to differentiate it from other sleep disorders causing excessive daytime sleepiness [22, 41]. PSG is also used and samples of cerebrospinal fluid can be taken to support a narcolepsy diagnosis [41, 68]. The treatment of these diseases usually focuses on easing the daytime sleepiness with pharmacological substances [68].

Parasomnias manifest in abnormal, unpleasant, or undesirable activities, behaviours, or experiences during sleep, at the onset of sleep, or during arousals from sleep [22, 41]. Moreover, parasomnias encompass NREM-related parasomnias (e.g. sleepwalking and sleep terrors), REM-related parasomnias (e.g. REM sleep behaviour disorder and nightmare disorder), and other parasomnias (e.g. sleep-related hallucinations) [41]. Parasomnias are occasionally associated with violent and disruptive behaviour and can often result in excessive daytime sleepiness and have been implicated in many psychiatric and neurological conditions [22]. When diagnosing parasomnias, a PSG with a video recording is often used whereas sleep diaries or home video recordings are sometimes sufficient [70]. The treatment of parasomnias initially focuses on inhibiting the potential for sleep-related injuries. Furthermore, parasomnias are occasionally related to other sleep disorders and therefore the treatment of these comorbid diseases may also ease parasomnia symptoms [71].

Sleep-related movement disorders manifest as involuntary movements during sleep [41]. These include disorders such as restless legs syndrome and sleep bruxism [41]. While some of these disorders are characterized by benign, unharmed movements causing no significant long-term consequences and generally resolving spontaneously, some may cause physical injury and long term damage (e.g. tooth wear and headache related to sleep bruxism) [41, 72, 73]. In the diagnosis of sleep-related movement disorders, it is crucial to be able to differentiate sleep-related movement disorders from other sleep disorders, especially from parasomnias. Thus, a PSG with a video recording is required in many instances while occasionally actigraphy, sleep diary, or EMG recording with audio may be sufficient [74, 75]. Treatment approaches are chosen according to the specific disorder; for example, sleep bruxism treatment focuses on preventing tooth wear while the treatment of restless legs syndrome relies on behavioural therapy and an improvement of sleep hygiene or on pharmacological substances [75]. Furthermore, sleep bruxism may occur alongside OSA in which case oral appliances used to prevent tooth wear are not applicable as they could further disrupt breathing [75].

In conclusion, while other diagnostic measurements and questionnaires exist, in-laboratory PSG is the most extensive method in sleep disorder diagnostics and remains the gold standard. Moreover, sleep staging from signals measured during a PSG forms the cornerstone of diagnosis.

2.3 BIOSIGNAL RECORDINGS

2.3.1 Polysomnography

The cornerstone of diagnosing sleep disorders is PSG utilizing a comprehensive measurement setup. To conduct the sleep staging, the electrical activity of the brain is measured via EEG, the eye movements via EOG, and the muscle tone via chin EMG. Additionally, a PSG includes the recording of photoplethysmogram (PPG) via a finger pulse oximeter which is also used to derive the blood oxygen saturation. Moreover, PSG commonly includes recordings used to assess respiratory effort (respiratory inductance plethysmography of thorax and abdomen), airflow (thermocouple, thermistor, and nasal pressure transducer), cardiac activity (electrocardiography, ECG), sleeping position (accelerometers or gravitation sensitive switches), the activity of the skeletal muscles in the legs (EMG), snoring sound (microphone or piezoelectric sensors). Finally, PSG also usually includes a video recording of the whole night [11,76].

Electroencephalography

The EEG is used to record the electrical activity of the brain. In PSG, EEG is measured noninvasively using multiple electrodes positioned on the scalp to identify the synchronous electrical potential over numerous neurons [77]. While EEG may lack the spatial resolution of the imaging techniques such as functional magnetic resonance imaging, it has superior temporal resolution making it ideal for sleep staging [78]. In PSG, EEG is measured using the frontal (F4-M1), central (C4-M1), and occipital (O2-M1) derivations with the placement conducted according to the International 10-20 System [79] (Figure 2.1). These are considered the minimum required channels but usually, backup electrodes (F4, C3, O1, and M2) are additionally used to provide substitutes in case of electrode malfunction [11].

The origin of the EEG signal lies in the synaptic activity of the neurons in the cerebral cortex. Each synaptic activity generates a small electrical impulse called the postsynaptic potential [77]. It is impossible to detect the postsynaptic potential of a single neuron with measurements conducted on the scalp; however, the synchronous postsynaptic potential of numerous neurons generates an electric field that can be measured with electrodes placed on the scalp [80]. The measurable potential is relatively small i.e. in the magnitude of microvolts; thus, an amplifier is required for signal collection [81].

The EEG signal from an electrode is represented as the difference in electrical potential to a reference electrode. A ground electrode can additionally be used for signal processing, for example, to prevent power line noise and amplifier drift [80]. Moreover, the electrodes typically demand the application of an electrolyte gel between the electrode and the skin. This is required as the electrochemical properties of the electrode-gel and gel-skin junctions lead to a steady electrical potential and impedance between the measured tissue and the measurement device [81].

Typically, EEG is recorded with a sampling frequency between 200 and 1000 Hz but even up to 5000 Hz frequencies can be used depending on the studied features [82]. In sleep studies, the minimum required sampling frequency is 200 Hz but an over 500 Hz sampling frequency is recommended [11]. After amplification, analog filtering can be implemented; however, digital filtering is often preferred to avoid losing any raw data [83]. In sleep studies, a high-pass filter with a 0.3 Hz and a low-pass filter with a 35 Hz cut-off frequency are recommended [11].

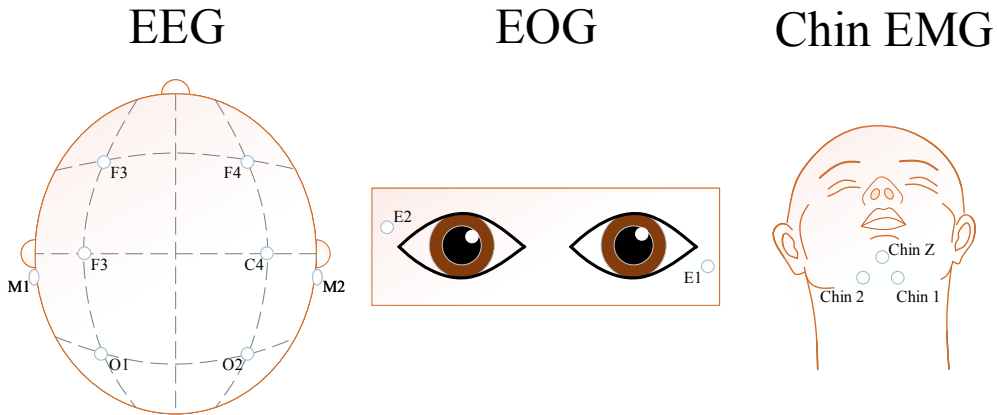


Figure 2.1: The electrode placement used in sleep staging for electroencephalography (EEG), electrooculography (EOG), and chin electromyography (EMG).

Electrooculography

EOG is used to capture eye movements. The basic principle behind EOG is that the eye can be considered as a dipole with a positive pole at the cornea and negative pole at the retina leading to a steady electric potential field. With eye movements, the position of the negative and positive poles change. In other words, the retina moves closer to one electrode while the cornea moves to the opposing electrode. Thus, the orientation of the dipole changes causing an alteration in the potential field leading to a measurable EOG signal [84]. For sleep staging purposes, the EOG is recorded with electrodes placed 1cm lateral and above the outer canthus of the left eye and 1 cm lateral and below the outer canthus of the right eye (Figure 2.1). Both are then referenced to the M2 electrode. This electrode positioning results in the out-of-phase deflections in the EOG with conjugate eye movements. In sleep studies, EOG is recorded with a minimum sampling frequency of 200 Hz but 500 Hz is recommended. Similarly to EEG, band-pass filtering between 0.3 Hz and 35 Hz is recommended. During sleep staging, the EOG measurement can be used to detect both slow and rapid eye movements [11].

Electromyography

Submental EMG is recorded to assess the electrical potential generated by muscle cells in the chin [76]. The chin EMG recording setup used for sleep staging consists of three electrodes. One electrode is placed 1 cm above the inferior edge of the mandible in the midline. Two electrodes are then placed symmetrically to 2 cm on the left and right sides. Moreover, the position of these two electrodes is 2 cm below the inferior edge of the mandible (Figure 2.1). The electrodes on the sides are referenced to the electrode in the middle. Either one of the derived channels is then used in sleep staging to assess the muscle tone. An EMG is required when identifying REM sleep. Similarly to EEG and EOG, the minimum sampling frequency is 200 Hz with 500 Hz as the recommendation. The EMG is filtered with 10 Hz high-pass and 100 Hz low-pass filters [11].

Photoplethysmography

Photoplethysmography (PPG) measures blood volume changes in tissue via optical sensors and light sources in a pulse oximeter [85]. The basic principle behind PPG is the absorption of light in hemoglobin. Generally, the other tissue components reflecting and scattering light do not vary in time. Thus, the absorption depends on changes in the blood volume within the measured tissue [86].

There are two ways to measure PPG: transmissive and reflective. Transmissive PPG is measured by placing the light detector directly across the light source and measuring the intensity of the transmitted light through the tissue. In reflective PPG, the light detector is placed near the light source to measure the intensity of back-scattered light [85, 86]. Typically, a pulse oximeter employs two different wavelengths: one infrared at around 940 nm and one red with around 660 nm wavelength. The infrared light provides a more stable signal over time whereas the red is more sensitive to changes in the oxygen concentration bound to hemoglobin in the blood volume [85]. Typically, the PPG signal produced by commercial pulse oximeters is the one formed with infrared light and it is typically heavily preprocessed in the hardware [85]. The blood oxygen saturation can be derived from the PPG as oxyhemoglobin absorbs less red and more infrared light whereas deoxyhemoglobin absorbs more red and less infrared light [86].

During sleep studies, the PPG signal is mainly used to derive the blood oxygen saturation; for example, this is of critical importance in diagnosing OSA. However, PPG contains a plethora of other information that has mainly been neglected after discovering its properties in deriving blood oxygen saturation [85]. Aside from providing a way to estimate the heart rate via changes in the blood volume caused by arterial pulsations, PPG reflects the autonomic activity [87, 88]. Moreover, the declines in the pulse wave amplitude in PPG are correlated to cortical activity during sleep. Variations in the spectral components of EEG during arousals from sleep are also measurable in PPG [87].

2.3.2 Portable sleep monitoring

While PSG is considered as the most comprehensive method to diagnose sleep disorders and is used as the gold standard reference method to assess sleep, it suffers from its high cost, the large amount of manual work required, and its limited availability [14]. Moreover, PSG can have a negative impact on sleep quality due to sleeping in an unfamiliar environment with multiple electrodes and sensors attached [89]. Therefore, portable, unattended monitoring devices are also often used to conduct recordings in a home environment. These ambulatory polygraphies (PG) are mainly used in OSA diagnostics and are thus also called respiratory polygraphies or home sleep apnea tests (HSAT). In some healthcare systems, mainly in Europe, these are even the preferred diagnostic method over the in-laboratory PSG [51].

The main difference between PG and PSG is that PG lacks the recording of EEG. As PG is mostly used in diagnosing OSA, it commonly includes a PPG recording used to determine the oxygen saturation. In addition, the recordings of airflow, respiratory effort, and ECG are also often included. Other signals such as audio, leg EMG, and body position may also be recorded but these vary between manufacturers [14, 90].

While PG can be used for a reasonably accurate diagnosis of OSA, especially when the pre-test probability is high [91], the lack of EEG recording is the most significant limitation. Since the EEG is not recorded, currently the sleep architecture cannot be assessed in any meaningful way. This also prohibits an identification of arousals from sleep and the total sleep time cannot be defined. In OSA diagnosis, this manifests in missing all the arousal-related hypopneas and the inability to determine the total sleep time. These shortcomings cause the determined AHI values to differ significantly from those based on PSG [52,54,92]. Nonetheless, there have been developments in ambulatory systems recording EEG with self-applicable electrode sets enabling inexpensive and simple recording of EEG in a home environment [93–97]. However, these are not yet widely used clinically.

2.3.3 Actigraphy

Actigraphy relies on a small wrist-worn device monitoring movements based on an accelerometer. The main advantage of actigraphy over a PSG is the simplicity and the capability to easily monitor over extended periods. However, actigraphy only provides an estimate for the sleep/wake patterns and cannot provide insights into the sleep architecture. Actigraphy is currently the preferred method for the long-term monitoring of sleep and assessing sleeping behaviours and sleep hygiene. It is especially useful when diagnosing circadian rhythm sleep-wake disorders, insomnia, or hypersomnias [98,99]. In addition to failing to assess sleep architecture or arousals from sleep, actigraphy has low specificity and tends to significantly overestimate sleep duration in situations when the individual is lying still but awake in bed [98,100,101]. Therefore, while useful for many purposes, actigraphy fails to assist in diagnosing those sleep disorders requiring a more accurate representation of sleep.

3 Deep learning

Deep learning is a class of techniques and methods belonging to the broader context of machine learning and artificial intelligence [18]. Deep learning relies on artificial neural networks that take their inspiration from the functions and information processing of neural systems. In traditional programming and problem-solving, the goal is generally to state the solution explicitly based on a set of rules and processes. In contrast, the goal of deep learning is to develop systems and computational architectures that can adapt and learn directly from observational data and information [102].

Deep learning algorithms can roughly be divided into three categories: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning relies on a set of training data that includes an input with a set of labels. The goal of supervised learning is to devise an algorithm capable of learning the features in the input data associated with each label [18, 19]. Examples of supervised learning include classification of images or identification of sleep stages from signals such as EEG. Conversely, unsupervised learning does not employ labels or targets but the goal is to independently learn useful properties and structure in the dataset, for example, group variance [19]. These algorithms include the cluster analysis of data. In reinforcement learning, the aim is to develop a software agent learning and performing a task in an environment based solely on trial and error, without any external guidance [19, 103]. The agent learns to function in the environment in order to maximize the notion of cumulative reward or minimize the penalty [103]. Reinforcement learning is often used in robotics and in tasks such as learning to play games.

In the following sections, a brief overview of the basic concepts of feedforward neural networks in the context of supervised learning are provided. Subsequently, the two main components of deep learning utilized in this thesis, convolutional neural networks and recurrent neural networks, are presented.

3.1 FULLY CONNECTED FEEDFORWARD NEURAL NETWORKS

The goal of a fully connected feedforward neural network is to learn how to approximate an arbitrary function f^* based on a set of data. Therefore, a neural network defines a mapping

$$\mathbf{y} = f(\mathbf{x}; \boldsymbol{\theta}), \quad (3.1)$$

where $\boldsymbol{\theta}$ represents the parameter values. The network learns and optimizes these parameters in order to achieve the best approximation of the function f connecting the input \mathbf{x} to the output \mathbf{y} [19]. For example, the goal of a classification task (e.g. identifying sleep stages) is to map an input \mathbf{x} (e.g. EEG signal segments) to a category \mathbf{y} (e.g. sleep stage). Moreover, the networks are generally represented by

concatenating numerous functions in a chain structure

$$f(\mathbf{x}) = f^{(n)} \circ f^{(n-1)} \circ \dots \circ f^{(2)} \circ f^{(1)}(\mathbf{x}), \quad (3.2)$$

where $f^{(i)}$ is the i th layer of the network and n is the number of layers that define the depth of the network. Moreover, the first layer must match the dimensionality of the input data while the final layer of the network is the output layer, providing the class label in classification problems. The layers between the first and last layer are generally called hidden layers [19]. The function compositions form a network with a certain depth leading to the terms "neural network" and "deep learning" being used. The definition of when a neural network can be considered as a deep neural network or as deep learning is somewhat vague. Historically, neural networks comprised a single hidden layer; thus, neural networks with more than two hidden layers are often considered as deep neural networks but more layers are commonly used [102].

In the equation (3.2), each hidden layer in the network defines a mapping $f^{(i)} : \mathbb{R}^n \rightarrow \mathbb{R}^m$. However, instead of interpreting a single layer as a vector-to-vector operation, these can be interpreted as m number of parallel units (neurons) each forming a $\sigma : \mathbb{R}^n \rightarrow \mathbb{R}$ vector-to-scalar function. That is, in a fully connected neural network, every neuron receives an input from all the neurons in the previous layer and operates on these to provide a single output, which is then passed to all the neurons in the following layer. This procedure is inspired by the biological function of neurons where the activation of a single neuron depends on the signals (electrical impulses) received from its multiple dendrites via synaptic connections of varying strengths to other neurons. Upon activation, the neuron outputs a signal along its single axon which eventually branches and connects to the dendrites of multiple other neurons. With neural networks, the activation a_i^l of a single neuron i in a layer l can be presented mathematically as

$$a_i^l = \sigma \left(\sum_j w_{i,j}^l x_j^{l-1} + b_i^l \right), \quad (3.3)$$

where σ is the nonlinear activation function, the sum is over all the single neurons j in the previous $(l-1)$ th layer, $w_{i,j}$ is the weight between the neuron i in the l th layer and neuron j in the $(l-1)$ th layer, x_j^{l-1} is the input originating from the neuron j in the previous layer, while b_i^l is the bias offset induced by the neuron [19,102]. As the input x to a neuron is defined by the activation of the neuron in the previous layer, the activation of a single layer of neurons can be further written in a more compact matrix form as [102]

$$\mathbf{a}^l = \sigma \left(\mathbf{w}^l \mathbf{a}^{l-1} + \mathbf{b}^l \right). \quad (3.4)$$

The final activation of the neural network is calculated by propagating the activation of all the layers to the following layers, similarly to the function compositions in equation (3.2).

In the most simplified form, activation of a single neuron can be presented by a perceptron [104]

$$a = \begin{cases} 0, & \text{if } \sum_{i=1}^n w_i x_i \leq b \\ 1, & \text{if } \sum_{i=1}^n w_i x_i > b. \end{cases} \quad (3.5)$$

However, perceptrons are seldom used due to their binary output [102]. More common nonlinear activation functions in the hidden layers include the hyperbolic

tangent (tanh) function and rectified linear unit (ReLU = $\max(0,x)$). In the final layer of classification problems, usually either the sigmoid function $\sigma_S(x) = 1/(1 + e^{-x})$ or softmax function is used. Of these, the softmax normalizes the output into a probability distribution illustrating the probabilities of each class in a single input [19].

The learning process of the neural network relies on the gradient-based minimization of a cost function. During training and after evaluating a single input with the neural network, a cost function maps the output of a model to a scalar value representing the difference between the output and the desired outcome, e.g., the class label. One possibility to define the cost for a single input x_k is with the mean squared error which can be written as

$$C_k(\boldsymbol{\theta}) = \frac{1}{2} \|y_k - f(x_k; \boldsymbol{\theta})\|^2, \quad (3.6)$$

where y_k is the desired output of the network and $f(x_k; \boldsymbol{\theta})$ is the output of the neural network with the input data x and internal parameters $\boldsymbol{\theta}$ [19, 102]. The cost function for the whole neural network then becomes

$$C_{\text{MSE}}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{k=1}^N C_k(\boldsymbol{\theta}), \quad (3.7)$$

where N is the total number of training inputs [102]. With this notation, the aim and learning process of the network become clear: the goal is to modify the parameters $\boldsymbol{\theta}$ (weights and biases related to neurons) such that the cost function $C(\boldsymbol{\theta})$ becomes minimized. However, it must be noted that the mean squared error may not be best suited to classification problems which deal with a set of known class labels and can suffer from the learning slowing down. Instead, cross-entropy may be often better suited, and can be presented as [102]

$$C_{\text{CE}}(\boldsymbol{\theta}) = -\frac{1}{N} \sum_{k=1}^N [y_k \ln f(x_k; \boldsymbol{\theta}) + (1 - y_k) \ln(1 - f(x_k; \boldsymbol{\theta}))]. \quad (3.8)$$

The main advantage of cross-entropy is that it applies a larger loss whenever there is a large difference between the desired output y and the predicted value $f(x; \boldsymbol{\theta})$. Thus, it is capable of better handling the slowing down of learning compared to mean square error in classification problems.

After the forward pass where the input is fed through the neural network and the value of the cost function is calculated, back-propagation is implemented. Back-propagation is used to allow the information of the cost to flow backwards through the network and to compute the gradient which is then used to update the parameters $\boldsymbol{\theta}$. The actual learning process is then conducted by changing the network parameters $\boldsymbol{\theta}$ to minimize the cost. This is done by changing the parameter values towards the negative gradient of the cost function [19, 102]. This can be implemented using algorithms such as stochastic gradient descent (SGD). Calculating the cost and changing the parameters after each input has been propagated through the network is often impractical and tends to halt the training process at local minima instead of the global minimum [19]. Therefore, in SGD, the input data is fed into the network in batches and the overall cost of the batch is calculated and back-propagated. Moreover, updating the parameters of the network occurs by taking a small, predefined step towards the negative gradient. This step

size is defined as the learning rate of the neural network. Updating the parameters of the network is then conducted as

$$\theta_{i+1} = \theta_i - \alpha \nabla C(\theta_i), \quad (3.9)$$

where θ_i are the parameter values at propagation i , α is the learning rate, and $C(\cdot)$ is the chosen cost function. Moreover, to avoid a termination of the learning at the local minima of the cost function, the SGD can be further supplemented with a momentum term or with adaptive algorithms such as RMSProp or Adam [19].

When developing neural networks, the dataset is generally divided into training, validation, and test sets. The training set is used for the actual learning process, meaning that it is used both for forward pass and back-propagation. The validation set is used to evaluate the performance of the network during training. After each batch of the training set is fed into the network and the parameters are optimized, the validation set is passed forward through the network and the value of the cost function is calculated [102]. However, the back-propagation process is not conducted for the validation set and it is not used to optimize the parameters. The validation set provides a measure of the network performance during training and can be used to choose the optimal model and avoid overfitting the network to the training data. Finally, the test set is used only once to measure the performance and generalisability of the final deep learning model. Occasionally, and especially with smaller datasets, k -fold cross-validation is also used. In this process, the dataset is partitioned into k subsets and one subset is used as the validation set with the remaining subsets utilized as the training set. The training process is then repeated k times until all the partitions have been used once as the validation set [19].

3.1.1 Convolutional neural networks

In contrast to the fully-connected neural networks described in the previous section, convolutional neural networks (CNNs) replace the matrix multiplication between the weights and inputs in at least one of the layers with a convolution operation. Therefore, it is specialized for data with a known grid-like topology, such as 1D time-series or images with a 2D grid of pixel values. The biological inspiration behind CNNs is the function of the visual cortex; this brain region functions based on a spatial map where neurons only respond to stimuli in a certain region, with these regions overlapping between multiple neurons [19].

CNNs are commonly based on four different types of layers: convolutional layers, activation layers, pooling layers, and fully connected layers. The main operation performed in a convolutional layer is a discrete convolution between two matrices: the input matrix I and the kernel matrix K of the layer. The kernel comprises the parameters that are changed during the training. In a generalized form, the discrete convolution can be defined as

$$(f * g)(i) = \sum_{a=-\infty}^{\infty} f(a)g(i-a), \quad (3.10)$$

where f and g are functions defined only at discrete integer values i [19]. In convolutional networks, the convolutions can be conducted with different dimensionalities of the input data, for example, 1D in the case of time series, and 2D with images. In these cases, the convolution between the input matrix I and kernel

K can be written as

$$S(i) = (K * I)(i) = \sum_m I(i - m)K(m), \quad (3.11)$$

in the 1D case, and

$$S(i, j) = (K * I)(i, j) = \sum_m \sum_n I(i - m, j - n)K(m, n), \quad (3.12)$$

in the 2D case. From these expressions, the convolution operator can also be interpreted as sliding the kernel across the whole input and calculating the product or cross product with the input in each position. Moreover, to ease the computational load, it is possible to skip some positions when moving the filter. This number of positions skipped is called stride and implementing it will both reduce the computational load and downsample the output of the convolutional layer [19].

After each convolutional layer, an activation layer is required. These function in the same way as occurs in fully connected feedforward networks and the same activation functions are applicable. Many CNN architectures include additional pooling layers after some of the activation layers. Pooling layers are used to downsample the output of the previous layer by calculating summary statistics of nearby outputs. Max-pooling is often used and outputs the maximum value of adjacent outputs [19]. Similarly, global average pooling can be used, but instead of outputting statistics from adjacent outputs, it outputs the average from over one whole data dimension. Global average pooling is often applied to replace the fully connected layers used to generate the output after the complete CNN architecture [105].

One of the main advantages of CNNs is the ability to combine the information of adjacent input values. For example, with time-series data, the adjacent values in the signal are usually connected. Similarly with 2D images, the adjacent pixel values are usually related to one another. Fully connected neural networks rely on matrix multiplication operations between neurons in adjacent layers, meaning that every individual neuron in a layer interacts with every unit in the preceding and following layers. CNNs, however, enable sparse connectivity, leading to better computational efficiency while still detecting meaningful features in smaller subsets of the complete input to the layer [19]. This further enables the implementation of deeper networks than can be achieved with fully-connected layers. CNNs have provided state-of-the-art results in many complex tasks, for example, they have surpassed previous image classification algorithms [18].

3.1.2 Recurrent neural networks

Recurrent neural networks (RNNs) are specialized in handling sequential data. While CNNs process data with a grid-like topology, RNNs process a sequence of data, for example, a sequence of 1D values in the case of a time-series or a sequence of 2D images from a video recording. The advantage of RNNs over CNNs is their scalability to longer sequences and their capability to natively handle sequences of variable lengths.

In contrast to feedforward networks where the input only flows forward, RNNs enable recurrent feedback connections. In its simplified form, the state \mathbf{h} of the hidden units in the RNN can be written as a dynamical system

$$\mathbf{h}^{(t)} = f\left(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}; \boldsymbol{\theta}\right), \quad (3.13)$$

where the state at t depends also on the previous states [19]. Additionally, RNNs usually incorporate output layers using the states of the hidden units to make predictions based on the input.

One of the most effective types of RNN is called gated RNN which relies on creating paths for the gradient to flow during back-propagation. These are aimed at more efficient learning and avoiding the issue of a vanishing gradient [19]. The most commonly used gated RNNs are the long short-term memory (LSTM) [106] and gated recurrent unit (GRU) networks [107].

LSTM networks are based on LSTM cells that have both recurrent connections between cells, but also self-loops to enable gradient flow. Aside from having similar inputs and outputs for each cell as in regular RNNs, LSTMs introduce gating units to control the information flow. The main components of the LSTM network are the forget gate, the external input gate, and the output gate. All these gates have their own parameters and weights and affect the output of the network. The forget gate unit $f_i^{(t)}$ is used to control the self-loop weight, and can be written for cell i at time step t in the form of

$$f_i^{(t)} = \sigma\left(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)}\right), \quad (3.14)$$

where σ is often chosen as the sigmoid function, \mathbf{h} contains the hidden unit states and all the outputs of the cells at the time step and \mathbf{b}^f contains the bias terms, \mathbf{U}^f the input weights, and \mathbf{W}^f the recurrent weights for the forget gates [19]. Similarly, the external input gate $g_i^{(t)}$ has its own parameters for each cell and can be written as

$$g_i^{(t)} = \sigma\left(b_i^g + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g h_j^{(t-1)}\right). \quad (3.15)$$

The output gate $q_i^{(t)}$ can similarly be written as [19]

$$q_i^{(t)} = \sigma\left(b_i^q + \sum_j U_{i,j}^q x_j^{(t)} + \sum_j W_{i,j}^q h_j^{(t-1)}\right). \quad (3.16)$$

Due to the sigmoid function σ , all the gates obtain values between 0 and 1 and determine the weights affecting the state of the cells and output of the network. The forget gate $f_i^{(t)}$ and the external input gate $g_i^{(t)}$ determine the internal state of the LSTM cell. The forget gate sets the weight for the previous state of the LSTM cell, while the external input gate sets the weight for all of the external input. With these, the internal state $s_i^{(t)}$ of a single LSTM cell can be written as

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \sigma\left(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)}\right), \quad (3.17)$$

where \mathbf{b} , \mathbf{U} , and \mathbf{W} are the biases, input weights, and recurrent weights of the LSTM cell, respectively [19]. The output $h_i^{(t)}$ of the LSTM cell depends on the state of the cell which is then regulated by the output gate $q_i^{(t)}$. With a hyperbolic tangent activation, the cell output can be written as [19]

$$h_i^{(t)} = \tanh \left(s_i^{(t)} \right) q_i^{(t)}. \quad (3.18)$$

In conclusion, a single LSTM cell has its own parameters (bias, input weights, and recurrent weights) and parameters for each of the gates (forget, external input, and output). The learning process is regulated by the gates and these enable the flow of the gradient across the network thus improving the training process. Finally, it should be noted that this chapter only describes a single version of the LSTM as other variations do exist.

GRUs are essentially similar to LSTMs but are designed to remove less important components of the LSTM. Therefore, the main difference between GRU and LSTM is that a GRU employs only a single gating unit that simultaneously controls the forgetting factor that determines the effect of previous states and the weight for updating the state unit [19]. The GRU cell only relies on an update gate $u_i^{(t)}$ and a reset gate $r^{(t)}$. The expressions for these gates are

$$u_i^{(t)} = \sigma \left(b_i^u + \sum_j U_{i,j}^u x_j^{(t)} + \sum_j W_{i,j}^u h_j^{(t-1)} \right), \quad (3.19)$$

and

$$r_i^{(t)} = \sigma \left(b_i^r + \sum_j U_{i,j}^r x_j^{(t)} + \sum_j W_{i,j}^r h_j^{(t-1)} \right). \quad (3.20)$$

These can then be used to write the updated equations for the GRU cells as [19]

$$h_i^{(t)} = u_i^{(t-1)} h_i^{(t-1)} + \left(1 - u_i^{(t-1)} \right) \sigma \left(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} r_j^{(t-1)} h_j^{(t-1)} \right). \quad (3.21)$$

Aside from enabling information from past values to affect the current state, it is also possible to develop RNNs that depend on the whole input sequence containing both the past and the future states. These RNNs are called bidirectional RNNs and combine two RNNs: one moving forward through time, starting from the beginning of the sequence and another moving in the opposite direction from end to start [19]. These can be especially useful for tasks such as speech recognition where the correct interpretation of a single word may depend on both previous and future contexts. Overall, RNNs have been exploited extremely successfully in tasks with long-term dependencies and sequential structure, such as natural language processing, handwriting recognition, and generating text or music [18,19].

3.2 APPLICATIONS IN MEDICINE

Deep learning applications in medicine have mainly focused on a more efficient and accurate diagnosis of medical conditions. Deep learning-based algorithms have especially excelled in analysing and interpreting medical images [108]. These methods have been able to reach the accuracy of expert dermatologists in diagnosing skin cancer and in differentiating between malignant and benign skin lesions [17]. Similarly, automatic identification and severity assessment of osteoarthritis from radiographs have been conducted based on a deep learning methodology [16]. Deep learning has also been successfully applied to automatically analyse biosignals. For example, deep learning has enabled ECG-based identification of arrhythmias, achieving a similar accuracy as a cardiologist [109] and even identifying patients with an elevated risk of developing ventricular dysfunction in the future [110].

Deep learning applications also exist in sleep medicine [111]. For example, machine learning methods have enabled an accurate determination of the AHI [112] and classification of OSA severity [113] from the oxygen saturation signal while a CNN-based analysis has been able to detect apnea and hypopnea events in a pediatric population [114]. Moreover, deep learning has been exploited to achieve an automatic diagnosis of type-1 narcolepsy with a high specificity [115]. A deep learning-based analysis of EEG signals has also been able to identify insomnia patients with a high accuracy [116] and unsupervised machine learning methodology has been applied to cluster OSA phenotypes based on PSG recordings [117].

Recently, there have been attempts made to conduct automatic sleep staging based on EEG, EOG, and chin EMG recordings with deep learning (Table 3.1). The more traditional approaches have relied on simple machine learning classification algorithms using a set of extracted features from the signals [118–120]; however, there has been more and more interest in developing approaches using deep learning [121–131]. These approaches have surpassed previously developed automatic methods. However, these deep learning approaches have also often relied on a heavy preprocessing of the input data and artificially transforming the 1-dimensional signals into 2-dimensional spectrograms [123, 126, 128, 130] or several predefined features extracted from the signals [129, 131] instead of utilizing the full potential of CNNs and RNNs in an end-to-end fashion. Additionally, the deep learning approaches have generally relied on research datasets of healthy individuals and have suffered from a low generalization to clinical populations [128]. Additionally, there have been promising deep learning-based sleep staging approaches utilizing only a single EEG-channel [121–125, 132, 133]. While these almost attain the accuracies of multi-channel approaches, they suffer from the same issues. Large clinical datasets collected during the normal clinical inflow of patients have rarely been used and the effect of sleep disorders on sleep staging has not been thoroughly investigated.

To avoid the laborious EEG recording that requires meticulous placement of electrodes, attempts have been made to exploit deep learning-based approaches to identify the sleep stages based on ECG [134–137]. These have usually relied on heart rate variability (HRV) characteristics [134, 137] combined with either respiratory effort [135] or actigraphy [136, 137]. Recently, there have also been attempts made to classify sleep stages based on photoplethysmogram (PPG) signals [138–142]. However, these have exploited more traditional machine learning classifiers and manual feature selection instead of relying on deep learning (Table 3.2) and

have mainly been based on substituting the ECG recording with PPG in the HRV estimation. Furthermore, the best performing approaches rely on additional actigraphy recording instead of relying solely on the PPG. Most of the approaches that have been attempted to assess sleep without EEG recording have focused on differentiating between wakefulness and sleep or identifying NREM and REM sleep. Thus, important parts of the sleep architecture remain overlooked, hindering their applicability to clinical populations. Moreover, the developed approaches have not reached a similar performance as manual analysis or EEG-based automatic methods.

Table 3.1: Summary of the most notable and successful previous approaches for automatic sleep staging based on deep learning.

	Method	Input	Dataset	Accuracy
Phan et al. [132]	Sequence-to-sequence RNN	Fourier transform of a single EEG channel	153 PSGs of healthy adults from Sleep-EDF	82.6%
Phan et al. [126]	Sequence-to-sequence RNN	Fourier transform of EEG, EOG, and chin EMG channels	200 PSGs of healthy adults from MASS	87.1%
Supratak et al. [122]	Combined CNN and LSTM	EEG referenced to EOG (MASS), or a single EEG channel (Sleep-EDF)	200 PSGs from MASS and 39 PSGs from Sleep-EDF	86.2% and 82.0%
Mousavi et al. [121]	Sequence-to-sequence CNN and RNN	Single EEG channel	153 PSGs of healthy adults from Sleep-EDF	80.0%
Patanaik et al. [128]	CNN	Fourier transform of EEG and EOG channels	Research and clinical PSGs of 459 individuals	72.1%-89.8%
Biswal et al. [130]	Combined CNN and RNN	Spectrogram of EEG and EMG channels	>10 000 clinical and research PSGs	87.6%

Sleep-EDF [143,144] and Montreal Archive of Sleep Studies (MASS) [145] are publicly available open-access datasets. CNN = convolutional neural network, RNN = recurrent neural network, LSTM = long short-term memory network, EEG = electroencephalography, EOG = electrooculography, EMG = electromyography, PSG = polysomnography. In Patanaik et al. [128] the clinical datasets included patients with Parkinson’s disease or with suspected sleep disorders. In Biswal et al. [130], the clinical dataset included patients with suspected sleep disorders

Table 3.2: Previous studies conducting automatic sleep stage classification based on photoplethysmogram (PPG).

	Method	Recordings	Dataset	2-stage accuracy	3-stage accuracy	4-stage accuracy
Fonseca et al. [138]	Linear discriminant classifier	PPG and accelerometer	165 healthy adults	92%	73%	59%
Beattie et al. [139]	Linear discriminant classifier	PPG and accelerometer	60 healthy adults	-	-	69%
Uçar et al. [140]	k-nearest neighbors classifier	PPG	10 adults with OSA	73%	-	-
Dehkori et al. [141]	Multivariate logistic regression	PPG	160 healthy children	77%	-	-
Motin et al. [142]	Support vector machine	PPG	5 patients with sleep-disordered breathing	72%	-	-

The 2-stage accuracy denotes the classification accuracy when differentiating between wakefulness and sleep. Similarly, 3-stage denotes wakefulness/NREM/REM and 4-stage denotes wakefulness/light sleep/deep sleep/REM.

4 Aims of the thesis

Diagnosis of sleep disorders relies on overnight polysomnographies with the inherent time-consuming manual analysis. For example, sleep stages are manually identified based on EEG, EOG, and EMG signals which all require meticulous placement of multiple electrodes and subsequently the manual analysis can take up to hours for even an experienced scorer. Moreover, sleep stages are identified in non-overlapping 30-second epochs which are a historical remnant from an era when the recordings were printed on paper for analysis. Thus, many transitions between sleep stages may be overlooked [13,146]. Furthermore, the manual analysis often relies on arbitrary rules. For example, OSA severity is classified based on AHI thresholds which lack a solid scientific foundation [55].

To resolve these issues, four specific aims were undertaken:

1. Optimize the AHI thresholds used for severity classification of OSA in polygraphic studies to better correspond to the risk of all-cause mortality.
2. Implement deep learning methods for EEG-based automatic sleep staging with a minimal number of recorded signals and study how OSA severity affects the accuracy of the automatic sleep staging.
3. Accurately determine the sleep stages without EEG using only a PPG signal recorded with a finger pulse oximeter.
4. Use deep learning methods to assess the sleep architecture in more detail beyond the traditional sleep staging approach and study how the sleep architecture varies between different OSA severity categories.

5 Methods

The studies included in this thesis utilized clinical datasets collected from individuals with a clinical suspicion of OSA. In study **I**, survival analysis was utilized to evaluate the optimal AHI thresholds for classification of OSA severity. In study **II**, deep learning methods were utilized to develop an approach for automatic sleep staging based on EEG and EOG recordings. In study **III**, a deep learning model was developed for identifying sleep stages from a PPG signal recorded with a finger pulse oximeter. In study **IV**, the deep learning-based sleep staging approach developed in study **II** was applied to analyse the sleep architecture with a better temporal resolution to enable a more detailed assessment of sleep fragmentation.

5.1 STUDY POPULATIONS AND MEASUREMENT DEVICES

Study **I** utilized a follow-up dataset of 1989 suspected OSA patients (Table 5.1) who had undergone polygraphic recordings in Kuopio University Hospital, Kuopio, Finland during 1993–2003 due to a clinical suspicion of OSA. The recordings were conducted using a custom-made ambulatory device recording four channels: airflow with an oronasal thermistor, arterial oxygen saturation with a pulse oximeter (Minolta Pulsox 7), sleeping position with tilt switches, and abdominal respiratory movements with a piezoelectric sensor [147]. All the signals were sampled with a 4 Hz sampling frequency. All the recordings were manually reanalysed during 2012–2015 [148] in compliance with the prevailing AASM criteria. Hypopneas were scored using a desaturation threshold of 4% (2007 AASM rule 4A) [11]. The background information of patients was collected from their medical records including BMI, smoking status, gender, age, CPAP treatment, and co-morbidities. Information on smoking and BMI was missing from a total of 206 patients who were excluded from the study population, leading to a total of 1783 patients being included. The causes of death were obtained from Statistics Finland (Helsinki, Finland) in February 2018. The mean (standard deviation) follow-up time was 18.3 (5.2) years. A favourable statement for retrospective data collection and analysis was obtained from The Research Ethics Committee of the Hospital District of Northern Savo, Kuopio, Finland (decision numbers 127/2004 and 24/2013).

Studies **II–IV** were based on a dataset of clinical PSGs of 933 patients with a clinical suspicion of OSA (Table 5.2). The recordings were conducted in the Princess Alexandra Hospital, Brisbane, Australia during 2015–2017 with the Compumedics Grael acquisition system (Compumedics, Abbotsford, Australia). Each recording was manually scored in compliance with the prevalent AASM guidelines [11] by one of ten experienced scorers participating regularly in intra- and inter-laboratory scoring concordance activities. The retrospective data collection was approved by the Institutional Human Research Ethics Committee of the Princess Alexandra Hospital (HREC/16/QPAH/021 and LNR/ 2019/QMS/54313). Studies **II** and **IV** utilized the EEG and EOG recordings sampled at 1024 Hz, the sleep stage scorings, and the identified respiratory events for determining the OSA severity. A total of 891 individuals had successful recordings of the required signals together with

Table 5.1: Demographic information of the patients in the follow-up dataset collected at the Kuopio University Hospital ($n = 1783$) and utilized in study I.

	<i>n</i> (% of the population)
Female	422 (24%)
Male	1361 (76%)
Non-smoker	766 (43%)
Quit smoking	480 (27%)
Smoker	537 (30%)
CPAP treatment	366 (21%)
Acute myocardial infarction	102 (6%)
Cardiovascular disease	231 (13%)
Diabetes	363 (20%)
	Median (Q₁ – Q₃)
Apnea-hypopnea index (h⁻¹)	5.7 (1.7 – 16.4)
Body mass index (kgm⁻²)	28.4 (25.5 – 32.7)
Age (years)	48.2 (41.3 – 55.1)

CPAP = continuous positive airway pressure, Q₁ = 25th percentile, and Q₃ = 75th percentile.

completed scorings and were included in these studies. Instead of EEG and EOG recordings, study III was based on PPG signals ($f_s = 256$ Hz), scored sleep stages, and scored respiratory events leading to a total study population of 894 individuals.

In study II, a publicly available dataset, Physionet Sleep-EDF [143, 144], was additionally used in addition to the clinical dataset presented above to enable comparison with previous studies. Version 2 of the expanded Sleep-EDF dataset released in March 2018 and comprising 153 PSGs was used. The dataset comprised 37 men and 41 women previously investigated to study the effect of age on sleep in a healthy population and most individuals had undergone two PSGs. From the recordings, the Fpz-Cz EEG signal and a horizontal EOG signal were used along with the manual sleep staging. The sleep staging was originally conducted according to the Rechtschaffen and Kales manual [12]. Each epoch was scored either into wake, one of four stages of NREM sleep, REM sleep, movement, or alternatively the epoch remained unscored if the scoring was technically impossible. The NREM stages comprising deep sleep (S3 and S4 [12]) were combined into a single N3 sleep stage to correspond to the current AASM guidelines [11] and the epochs that were not scored or were scored as movement were left out due to their small number. Furthermore, the original recordings included long periods of wakefulness before and after sleep. Thus, the recordings were truncated to contain only up to 30 minutes of wakefulness before and after sleep to obtain more reliable results and to enable reliable comparison to previous studies. The percentage of sleep stages in the Sleep-EDF dataset and the clinical PSG dataset is presented in Table 5.3.

Table 5.2: Demographic information of the suspected obstructive sleep apnea patients (OSA) patients who underwent clinical polysomnographies at the Princess Alexandra Hospital. This dataset was utilized in studies **II** and **IV** ($n = 891$) and in study **III** ($n = 894$).

	Studies II and IV	Study III
	<i>n</i> (% of the population)	
Non-OSA (AHI < 5)	152 (17%)	154 (17%)
Mild OSA ($5 \leq \text{AHI} < 15$)	278 (31%)	278 (31%)
Moderate OSA ($15 \leq \text{AHI} < 30$)	208 (23%)	209 (23%)
Severe OSA (AHI ≥ 30)	253 (28%)	253 (28%)
Female	398 (45%)	398 (45%)
Male	493 (55%)	496 (55%)
	Median ($Q_1 - Q_3$)	
Age (years)	55.8 (44.7 – 65.8)	55.9 (44.7 – 65.8)
Arousal index (h^{-1})	20.8 (14.0 – 31.4)	20.7 (13.9 – 31.4)
Apnea-hypopnea index (h^{-1})	15.8 (7.0 – 32.8)	15.8 (7.0 – 32.6)
Body mass index (kgm^{-2})	34.5 (29.4 – 40.4)	34.4 (29.4 – 40.4)
Sleep efficiency (%)	70.7 (57.9 – 82.0)	70.7 (58.1 – 81.9)
Sleep latency (min)	17.5 (9.0 - 34.5)	17.5 (9.0 - 34.5)
Total recording time (min)	442.5 (409.5 – 474.5)	442.3 (409.5 – 474.0)
Total sleep time (min)	308.8 (253.8 – 359.8)	308.8 (253.5 – 359.5)
Wake after sleep onset (min)	102.8 (61.3 – 150.0)	102.5 (61.0 – 149.5)

AHI = apnea-hypopnea index, Q_1 = 25th percentile, and Q_3 = 75th percentile. Sleep efficiency is the percentage of sleep from the total recorded time, sleep latency is the time spent awake before falling asleep, and wake after sleep onset is the duration of wakefulness during the night after falling asleep.

Table 5.3: The percentage of sleep stages in the Sleep-EDF dataset [143, 144] and the clinical dataset including the percentages for each obstructive sleep apnea (OSA) severity.

	Wake	N1	N2	N3	REM
<i>Studies II and IV</i>					
Sleep-EDF ($n = 153$)	34%	11%	35%	7%	13%
Clinical dataset ($n = 891$)	32%	9%	33%	13%	12%
Non-OSA ($n = 152$)	28%	6%	35%	17%	14%
Mild OSA ($n = 278$)	29%	7%	36%	15%	13%
Moderate OSA ($n = 208$)	32%	32%	33%	14%	12%
Severe OSA ($n = 253$)	39%	15%	28%	9%	9%
<i>Study III</i>					
Clinical dataset ($n = 894$)	32%	9%	33%	13%	12%

5.2 OPTIMIZING THE AHI THRESHOLDS USED FOR OSA SEVERITY CLASSIFICATION

In study I, the AHI thresholds used for OSA severity classification were optimized for polygraphic recordings with regards to the risk of all-cause mortality. The optimization was based on survival analysis and various threshold combinations were simulated when classifying the patients into four groups: non-OSA, mild OSA, moderate OSA, and severe OSA. The minimum threshold value for mild OSA was 1 h^{-1} and the maximum for severe OSA was 80 h^{-1} . All the threshold combinations with a minimum separation of 1 h^{-1} between the thresholds was simulated leading to a total of 79079 combinations.

Corresponding to each combination of threshold values, the risk of all-cause mortality for each OSA severity category was investigated using the Cox proportional hazards model [149, 150]. In the model, mortality was used as the studied event and the follow-up time was used as the time to the event. Censoring was applied to all the individuals still alive after the follow-up time (i.e. when the mortality information was obtained). Furthermore, the model was adjusted for age, BMI, cardiovascular disease, CPAP treatment, diabetes, gender, smoking, and the occurrence of acute myocardial infarction.

To control for overfitting when optimizing the thresholds to the study population, separate optimization and validation groups were used with controlled random sampling. Instead of randomizing the whole population, the randomization was done in two steps to include patients with all the adjusting variables in each group during each randomization. The randomized sampling was conducted as follows:

1. From the complete population, women, patients with a pre-existing diagnosis of diabetes or cardiovascular diseases, patients treated with CPAP, and patients who suffered an acute myocardial infarction were all grouped into a single group and this group was then randomized to equally sized optimization and validation groups.
2. The remaining population was randomized to the optimization and validation groups until both groups were evenly sized.

After the split into optimization and validation groups, the Cox proportional hazards model was used to evaluate the risk of all-cause mortality with the different threshold combinations. This complete protocol was repeated 100 times, leading to a total of 7 907 900 formations of the proportional hazards model.

After simulating every threshold combination and after all the randomizations, the optimization criteria presented in Table 5.4 were applied to choose the optimal threshold combination.

Corresponding to each randomization, every threshold combination that fulfilled all of the optimization criteria 1-3 was selected. Out of all the randomizations, the threshold combination fulfilling all the criteria most often in the optimization set was chosen as the optimal threshold combination. This combination was then studied across all the validation sets and the median values of all the obtained hazard ratios were calculated.

Table 5.4: The criteria used for threshold optimization.

1.	The hazard ratios for all-cause mortality must increase nearly linearly when progressing towards more severe OSA to ensure that the more severe disease reflects a higher mortality risk. This criterion was applied by demanding that the differences in hazard ratios between mild and moderate and between moderate and severe OSA are within a margin of ± 0.02 .
2.	All the severity categories must include a minimum of 15% of the population.
3.	The number of patients in each severity category decreases with increasing OSA severity. That is, for group sizes it must hold that $n(\text{non-OSA}) > n(\text{mild OSA}) > n(\text{moderate OSA}) > n(\text{severe OSA})$.

5.3 DEEP LEARNING-BASED SLEEP STAGING

Studies **II** and **III** focused on developing deep learning approaches for automatic sleep staging. In study **II**, the EEG channel (derivation F4-M1) was used either alone for single-channel sleep staging or together with an EOG channel (derivation E1-M2) for a multi-channel approach. In contrast, study **III** utilized a PPG signal recorded with a transmissive finger pulse oximeter. Both studies were based on a combined convolutional (CNN) and recurrent neural networks (RNN). In both studies, the input signals were downsampled to 64 Hz to reduce the computational load. No additional preprocessing was required. The CNN architecture was chosen to learn the characteristic features of each sleep stage from the signals while the RNN was chosen to account for the temporal nature of sleep stages during the night. The implementation of the deep learning models was conducted in Python 3.6 using the Keras application programming interface (API) 2.2.4 with TensorFlow (v. 1.13) backend.

5.3.1 Neural network architecture

The CNN architectures in studies **II** and **III** were identical and comprised a total of six 1D convolutional layers, two max-pooling layers, and a global average pooling layer (Figure 5.1). Each convolution was followed by batch normalization and a rectified linear unit (ReLU) activation. The number of convolutional filters equalled the sampling frequency of the input signal for the first two convolutions, two times the sampling frequency for the next two convolutions, and four times the sampling frequency for the last two convolutions. The kernel size was 21 for the first two convolutions and 5 for the rest. The stride size was 5 for the first 1D convolution and 1 for the rest. The max-pooling layers were situated after the first two convolutions and after the next two and had a pool size of 2 with a stride size of 2. The global average pooling was the final layer of the CNN architecture.

The complete network architecture in studies **II** and **III** included a time distributed layer of the CNN, followed by a Gaussian dropout layer with rate of 0.3. In study **II**, these were then followed by a bidirectional LSTM layer with the number of units equalling four times the sampling frequency. In the forward step, a dropout rate of 0.3 and a tanh activation were used. In the recurrent step, a dropout rate of 0.5 and a hard sigmoid activation were used. In contrast, study **III** utilized a bidirectional gated recurrent unit (GRU) layer with the number of cells equalling four times the sampling frequency. The GRU was chosen due to its computational efficiency and a comparable performance to the LSTM [151,152]. A dropout of 0.3 was used in the forward step and 0.5 in the recurrent step. In both studies, the final layer of the model was a time distributed dense layer with softmax activation. This layer produced an output sequence of the sleep stage probabilities.

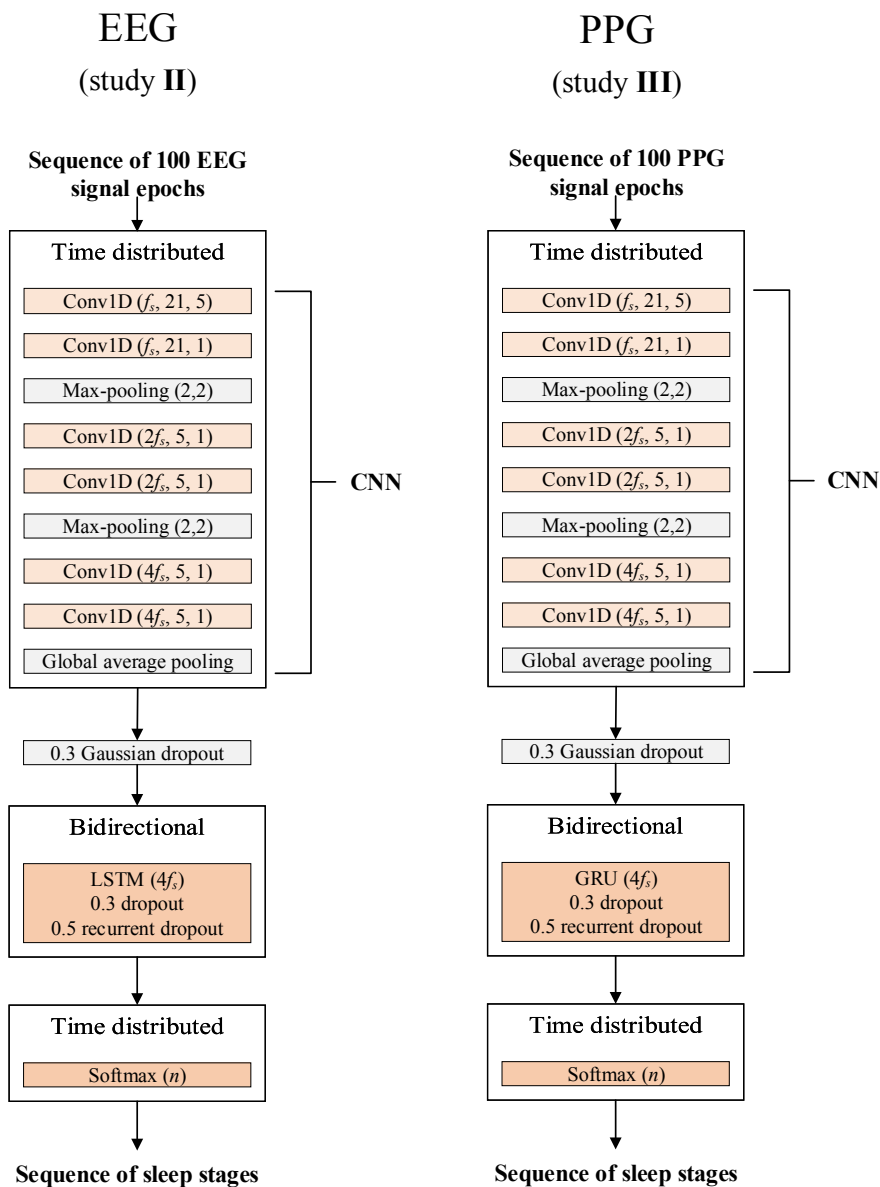


Figure 5.1: The architectures of the combined convolutional neural network (CNN) and recurrent neural network used in studies II and III. f_s denotes the sampling frequency of the input signal. The parameters are given for the 1D convolutions (Conv1D) as (number of filters, kernel size, stride size) and for the max-pooling as (pool size, stride size). The dropout rate is given for all the dropout layers and the number of units is given for the long short-term memory (LSTM) layer, gated recurrent unit (GRU) layer, and for the dense layer (Softmax). In study II, the input was a sequence of hundred 30-second epochs of either a single EEG channel or an EEG channel combined with an EOG channel. In study III, sequences of PPG signal were used as the input. The final output of the models was a sequence of softmax values representing the probability of each sleep stage for each epoch. n denotes the number of sleep stages used in the classification. This figure is reconstructed from studies II and III.

5.3.2 The training process and performance evaluation

In studies **II** and **III**, the model was trained using sequences of hundred 30-second epochs. During the training process, the size of the training data was multiplied fourfold by using an overlap of 75% when forming the sequences. However, the overlap was not used for validation or test sets. The training was conducted using a categorical cross-entropy loss function. Furthermore, an Adam optimizer with warm restarts [153] was used. The learning rate was optimized with a learning rate finder [154], and a range of 0.001 to 0.00001 was used in both studies.

EEG-based sleep staging

In study **II**, sleep staging was conducted in two different populations, in a public dataset (Sleep-EDF [143,144]) of healthy individuals and in a clinical dataset from Princess Alexandra Hospital. In the clinical dataset, the model was trained using the whole dataset and separately in each OSA severity category. The sleep staging was conducted either using a single EEG channel (F4-M1) or with a two-channel approach using an EEG (F4-M1) and EOG (E1-M2) channels.

In the Sleep-EDF dataset, ten-fold cross-validation was used to assess the performance of the model. During each fold in the cross-validation, 90% of the population was used for training and 10% as an independent test set. Furthermore, to avoid overfitting during training and to choose the optimal model, 10% of the training set was further chosen as the validation set during each fold. Ten-fold cross-validation was chosen due to the relatively small size of the dataset and to enable a comparison with the literature.

With the clinical dataset, the performance of the model was evaluated using the whole study population, with additional performance evaluation in each OSA severity category. When the complete study population was used, the dataset was split into three individual sets: a training set of 717 recordings (80%), a validation set of 87 recordings (10%), and an individual test set of 87 recordings (10%). To study the performance of the model in each OSA severity category, the dataset was split into the four groups (non-OSA, mild OSA, moderate OSA, and severe OSA) and the performance was evaluated in each group with ten-fold cross-validation using a similar procedure as with the Sleep-EDF dataset.

The accuracy of the sleep staging was calculated in an epoch-by-epoch manner. Furthermore, the correspondence of the deep learning-based sleep staging to manual staging was evaluated using Cohen's kappa coefficient (κ) [155]. The accuracy of each sleep stage was further evaluated by assessing the confusion matrix. When using ten-fold cross-validation, all the performance metrics were calculated over all the folds.

PPG-based sleep staging

In study **III**, the dataset comprising 894 individuals was split into a training set of 715 recordings (80%), a validation set of 89 recordings (10%), and an independent test set of 90 recordings (10%). The training process was conducted individually using three different sleep classification systems: 3-stage classification (wake/NREM/REM), 4-stage classification (wake/light sleep/deep sleep/REM), and 5-stage classification (wake/N1/N2/N3/REM).

The accuracy and κ were calculated and the confusion matrices formed in an epoch-by-epoch manner for each classification system. To further assess the usability of the PPG-based sleep staging, total sleep time and sleep efficiency were calculated. Furthermore, the values of the AHI were calculated corresponding to the manual sleep staging and when using the PPG-based approach. For further comparison, AHI calculation from polygraphic recordings was simulated by dividing the number of all the scored respiratory events by the total recording time. This index is occasionally also called the respiratory event index (REI) and is a common metric with polygraphic recordings [11].

5.4 DEEP LEARNING-BASED SLEEP STAGING WITH BETTER TEMPORAL RESOLUTION

In study **IV**, the deep learning model developed in study **II** was implemented to evaluate the sleep staging in more detail over the traditional approach with non-overlapping 30-second epochs. The model trained on the clinical population using the combination of EEG (F4-M1) and EOG (E1-M2) channels was used. Traditionally, the sleep stages are scored in non-overlapping 30-second epochs starting from the onset of the sleep study (Figure 5.2). In study **IV**, the deep learning model was used to score the sleep stages with the traditional approach and by allowing an overlap between consecutive 30-second epochs (Figure 5.2). Three different epoch-to-epoch durations were studied: starting a new 30-second epoch every 15 seconds (50% overlap), every 5 seconds (83.3% overlap), or every 1 second (96.7% overlap).

After scoring the complete study population with the four different approaches (traditional, 1-, 5-, and 15-second epoch-to-epoch duration), the individuals were grouped based on OSA severity categories. With each scoring, the sleep characteristics of the OSA severity groups were compared by calculating the sleep stage percentages, sleep parameters (i.e. total sleep time, TST; sleep efficiency, SE; and wake after sleep onset, WASO), and by studying the sleep fragmentation via survival analysis.

In the survival analysis, the event studied was awakening from sleep and the time to event was the mean duration of the sleep periods (i.e. onset of sleep until the next epoch scored as wake) for each individual. Thus, no censoring was needed for the individuals. When comparing the OSA severity groups, Cox proportional hazards model and Kaplan-Meier survival curves were used. The survival curves provided a graphical representation of the differences between groups and the Cox proportional hazards model provided the hazard ratios for each group illustrating the risk of fragmented sleep (i.e. short continuous sleep periods during the night).

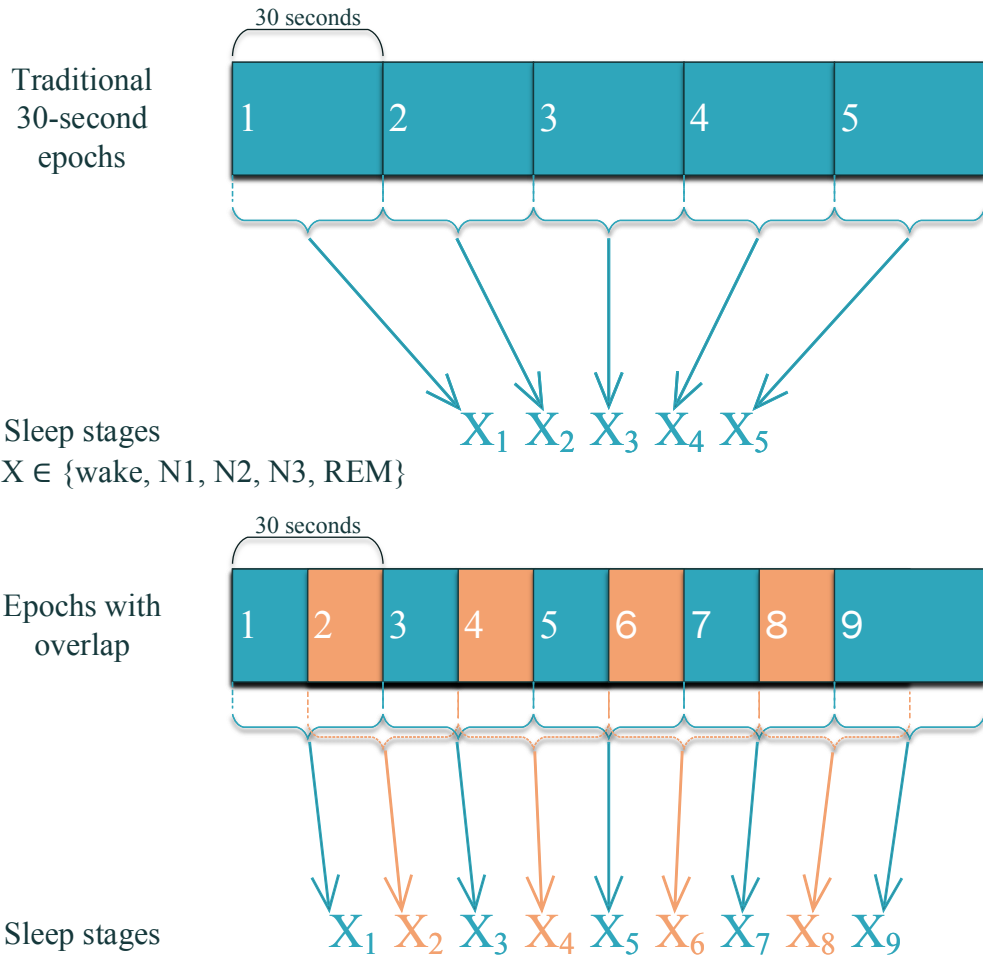


Figure 5.2: Sleep staging process with the deep learning model in study IV. The figure illustrates traditional sleep staging with 30-second epochs (top) and the more detailed approach by allowing the 30-second epochs to overlap (bottom). With the overlapping epochs, a sleep stage is identified for each epoch (X_i) which are then ordered based on the starting point of the epoch. X_i illustrates the identified sleep stage in epoch i , where $X \in \{\text{wake, N1, N2, N3, REM}\}$. In this figure, only the 15-second epoch-to-epoch duration is shown for clarity. Figure reconstructed from study IV.

5.5 STATISTICAL ANALYSES

In study I, the Cox proportional hazards model was used to evaluate the connection between OSA severity categories and all-cause mortality. In studies II and III, the inter-rater agreement between manual and deep learning-based sleep staging was assessed with Cohen's kappa (κ) coefficient [155]. Furthermore, in study III the statistical significance of the differences between sleep parameters (i.e. TST, SE, percentages of sleep stages, and AHI) derived from manual and automatic scoring was evaluated using the Wilcoxon signed-rank test. In study IV, the sleep continuity was determined based on the Cox proportional hazards model and Kaplan-Meier survival curves. Furthermore, the statistical significance of the differences in sleep parameters (i.e. TST, SE, WASO) between different scoring approaches was evaluated with the Wilcoxon signed-rank test and between the OSA severity categories with the Mann-Whitney U test. All statistical analyses were conducted with Matlab 2018b using the Statistics and Machine Learning Toolbox (The MathWorks, Natick, MA, USA) or in Python 3.6 using the scikit-learn library [156].

6 Results

The main result of study I was that the current AHI thresholds of 5-15-30 h⁻¹ used in the severity classification of OSA are not optimal and the combination of 3-9-24 h⁻¹ would better reflect the OSA-related risk of all-cause mortality. The main result of study II was the demonstration that a deep learning-based sleep staging could be achieved from a single frontal EEG channel with comparable accuracy to that of manual scoring. Study III showed that sleep staging can be conducted from a PPG signal measured with a simple finger pulse oximeter. In study IV, the use of overlapping 30-second epochs in the deep learning-based sleep staging enabled a more accurate representation of sleep architecture with better temporal resolution and revealed the highly fragmented sleep in patients with severe OSA better than possible with traditional non-overlapping 30-second epochs. The following sections describe the results of each study in more detail.

6.1 MORTALITY RISK-BASED AHI THRESHOLDS FOR OSA SEVERITY CLASSIFICATION

In study I, the AHI thresholds used for severity classification of OSA were optimized for ambulatory polygraphy with regards to the risk of all-cause mortality. The hazard ratios for the all-cause mortality risk varied greatly across all the simulated threshold combinations. The threshold dividing the non-OSA from mild OSA had the most observable effect on the magnitude of the hazard ratios (Figure 6.1). When this threshold was lowered from the current value of 5 h⁻¹, the hazard ratios for all the OSA severity categories increased. The thresholds between mild and moderate and between moderate and severe OSA exerted a smaller effect on the overall magnitude of the hazard ratios.

The threshold combination of 3-9-24 h⁻¹ fulfilled all of the optimization criteria (Table 5.4) most often (37 times) in the randomized optimization sets whereas the combinations 3-9-23 h⁻¹, 3-9-25 h⁻¹, and 3-9-26 h⁻¹ fulfilled the optimization criteria 34 times. An example of the threshold combinations fulfilling the optimization criteria during a single optimization group out of 100 randomizations is shown in Figure 6.2. With the optimized thresholds (3-9-24 h⁻¹), 630 patients were classified into the non-OSA category, 459 as mild, 377 as moderate, and 317 into the severe OSA category. In contrast, with the traditional thresholds (5-15-30 h⁻¹) 838 were classified into the non-OSA category, 469 as mild, 232 as moderate, and 244 into the severe OSA category.

Compared to the traditional thresholds, the optimized thresholds (3-9-24 h⁻¹) increased the hazard ratios in all of the OSA severity categories in the optimization datasets (Table 6.1) and the validation sets (Table 6.2). The median of the hazard ratios across the randomizations were 1.11 ($p = 0.50$), 1.61 ($p = 0.05$) and 1.64 ($p = 0.06$) for mild, moderate, and severe OSA, respectively, with the threshold combination of 5-15-30 h⁻¹ and increased to 1.41 ($p = 0.15$), 1.66 ($p = 0.05$) and 1.82 ($p = 0.03$), respectively, with the threshold combination of 3-9-24 h⁻¹.

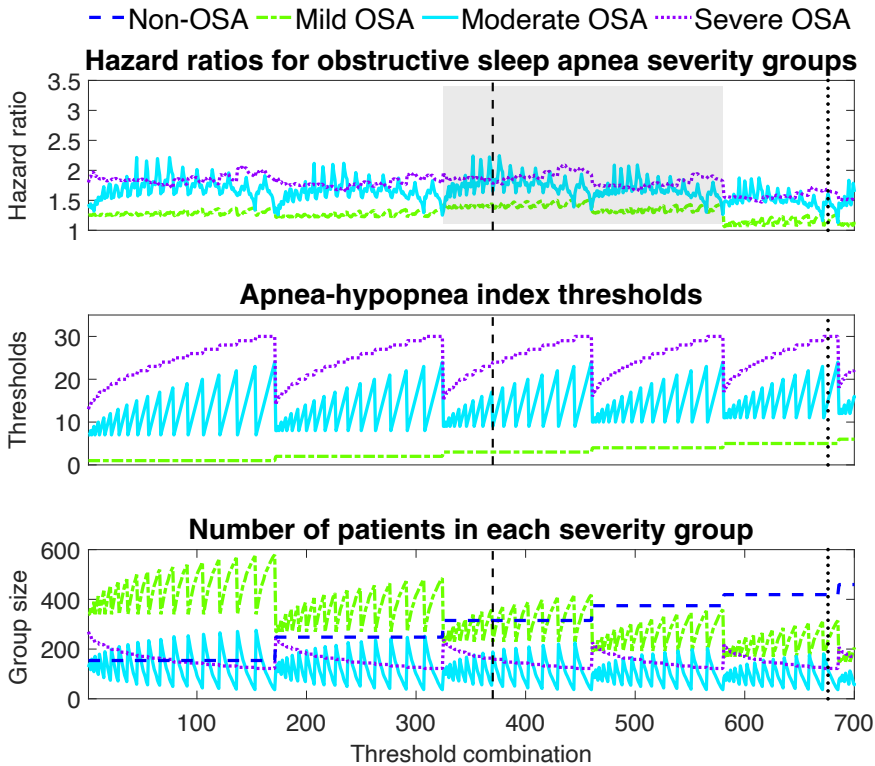


Figure 6.1: Illustration of the threshold simulations. The hazard ratios for all-cause mortality and the number of patients in each obstructive sleep apnea (OSA) severity category are shown corresponding to each threshold combination. Only a single optimization group out of 100 randomizations and a limited number of combinations (700 out of 79079) are illustrated for clarity. The vertical dashed line represents the optimized thresholds (3-9-24 h^{-1}) and the dotted line the traditional thresholds (5-15-30 h^{-1}). The shaded area represents the optimal area with the highest hazard ratios when the lowest threshold is between 3 and 5 h^{-1} . Figure reconstructed based on study I.

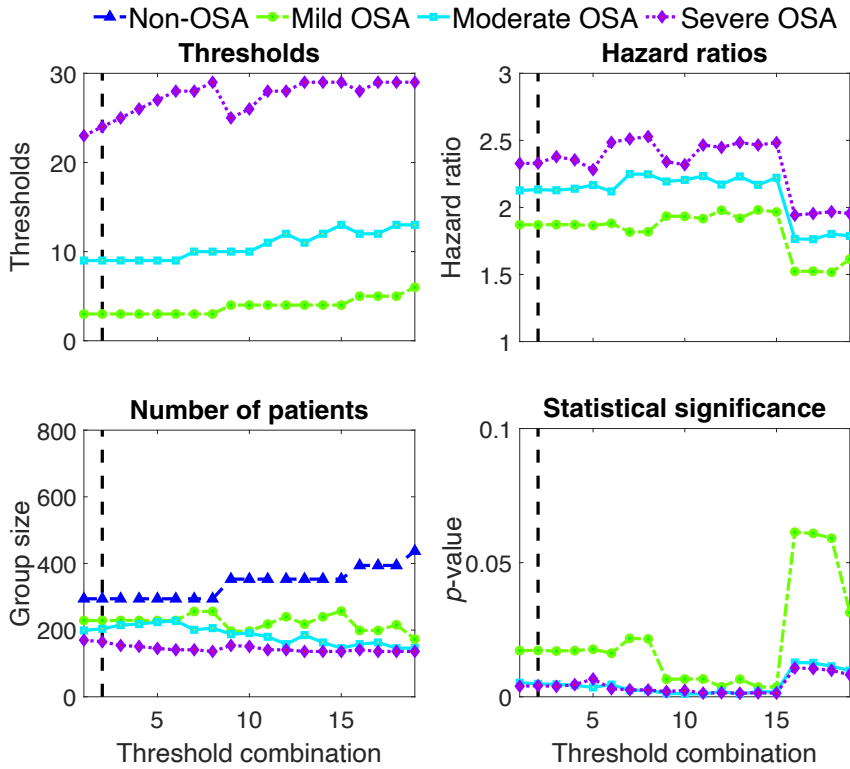


Figure 6.2: An example of the all threshold combinations fulfilling all the optimization criteria during a single optimization group out of 100 randomizations. The corresponding hazard ratios, number of patients, and the statistical significance are shown for each obstructive sleep apnea (OSA) severity category corresponding to each threshold combination. The optimal threshold combination of 3-9-24 h⁻¹ is indicated with a vertical dashed line. Figure reconstructed based on study I.

Table 6.1: The median hazard ratios and corresponding p -values calculated from all the 100 randomized optimization sets with the optimized and traditional severity classification thresholds.

	3-9-24 h ⁻¹				5-15-30 h ⁻¹			
	Hazard ratio		p -value		Hazard ratio		p -value	
	<i>Median</i>	<i>IQR</i>	<i>Median</i>	<i>IQR</i>	<i>Median</i>	<i>IQR</i>	<i>Median</i>	<i>IQR</i>
Mild OSA	1.43	0.31	0.14	0.25	1.13	0.22	0.56	0.45
Moderate OSA	1.64	0.39	0.04	0.09	1.64	0.38	0.03	0.10
Severe OSA	1.76	0.51	0.03	0.09	1.72	0.40	0.03	0.10
Age (risk/year)	1.08	0.01	<0.01	<0.01	1.08	0.01	<0.01	<0.01
AMI	1.55	0.36	0.05	0.14	1.56	0.38	0.05	0.14
BMI (risk/kgm⁻²)	1.03	0.01	0.06	0.06	1.03	0.01	0.02	0.08
CPAP	0.75	0.13	0.12	0.23	0.71	0.14	0.07	0.17
CVD	1.49	0.23	0.03	0.07	1.48	0.24	0.03	0.08
Diabetes	1.21	0.17	0.28	0.38	1.20	0.17	0.27	0.32
Male	1.44	0.33	0.08	0.20	1.49	0.32	0.07	0.16
Smoker	2.29	0.42	<0.01	<0.01	2.27	0.40	<0.01	<0.01
Former smoker	1.10	0.20	0.57	0.47	1.09	0.22	0.58	0.44

The hazard ratios for each obstructive sleep apnea (OSA) severity category were calculated using the Cox proportional hazards model adjusted for age, occurrence of acute myocardial infarction (AMI), body mass index (BMI), continuous positive airway pressure (CPAP) treatment, cardiovascular disease (CVD), diabetes, gender, and smoking status. IQR is the interquartile range calculated as the difference between 75th and 25th percentiles.

Table 6.2: The median hazard ratios and corresponding p -values calculated from all the 100 randomized validation sets with the optimized and traditional severity classification thresholds.

	3-9-24 h⁻¹				5-15-30 h⁻¹			
	Hazard ratio		p-value		Hazard ratio		p-value	
	<i>Median</i>	<i>IQR</i>	<i>Median</i>	<i>IQR</i>	<i>Median</i>	<i>IQR</i>	<i>Median</i>	<i>IQR</i>
Mild OSA	1.41	0.34	0.15	0.24	1.11	0.18	0.50	0.43
Moderate OSA	1.66	0.38	0.05	0.11	1.61	0.40	0.05	0.14
Severe OSA	1.82	0.51	0.03	0.13	1.64	0.41	0.06	0.14
Age (risk/year)	1.08	0.01	<0.01	<0.01	1.08	0.01	<0.01	<0.01
AMI	1.53	0.44	0.05	0.14	1.56	0.46	0.09	0.30
BMI (risk/kgm⁻²)	1.03	0.01	0.02	0.06	1.03	0.01	0.02	0.06
CPAP	0.78	0.16	0.21	0.39	0.73	0.16	0.14	0.32
CVD	1.49	0.32	0.04	0.12	1.49	0.30	0.05	0.14
Diabetes	1.19	0.20	0.34	0.43	1.22	0.20	0.28	0.41
Male	1.39	0.34	0.14	0.25	1.39	0.39	0.12	0.25
Smoker	2.34	0.52	<0.01	<0.01	2.33	0.51	<0.01	<0.01
Former smoker	1.10	0.20	0.53	0.45	1.10	0.22	0.53	0.47

The hazard ratios for each obstructive sleep apnea (OSA) severity category were calculated using the Cox proportional hazards model adjusted for age, occurrence of acute myocardial infarction (AMI), body mass index (BMI), continuous positive airway pressure (CPAP) treatment, cardiovascular disease (CVD), diabetes, gender, and smoking status. IQR is the interquartile range calculated as the difference between 75th and 25th percentiles.

6.2 DEEP LEARNING-BASED AUTOMATIC SLEEP STAGING BASED ON EEG AND EOG RECORDINGS

In study II, deep learning methods for automatic sleep staging based on single-channel (EEG) and two-channel (EEG + EOG) recordings were developed. The performance of the methods was evaluated on a public dataset of healthy individuals and in a clinical dataset of patients in whom there was a suspicion of OSA. Furthermore, the effect of OSA severity on the sleep staging accuracy was assessed.

6.2.1 Sleep staging in a public dataset of healthy individuals

In the updated Sleep-EDF dataset, the model achieved an epoch-by-epoch accuracy of 89.8% ($\kappa = 0.86$) in the training set, 83.0% ($\kappa = 0.77$) in the validation set, and 83.9% ($\kappa = 0.78$) in the test set with the two-channel input during the ten-fold cross-validation. With respect to the individual sleep stages, the accuracies were 93.7% for wake, 45.1% for N1, 87.3% for N2, 78.0% for N3, and 85.4% for REM in the test sets (Figure 6.3 A).

With the single-channel approach, the accuracies were 89.2% ($\kappa = 0.85$), 82.8% ($\kappa = 0.77$), and 83.7% ($\kappa = 0.77$) in the training, validation, and test sets, respectively. In the test sets, the individual sleep stage accuracies were 93.4% for wake, 43.4% for N1, 87.3% for N2, 78.7% for N3, and 85.4% for REM (Figure 6.3 B). In the updated Sleep-EDF dataset, the accuracies obtained surpassed previously published results (Table 6.3).

6.2.2 Sleep staging in a clinical dataset of patients with suspected OSA

In the clinical dataset of suspected OSA patients, the two-channel epoch-by-epoch accuracy was 85.5% ($\kappa = 0.80$) in the training set, 83.8% ($\kappa = 0.78$) in the validation set, and 83.8% ($\kappa = 0.78$) in the independent test set. With respect to the individual sleep stages, the accuracies were 89.4% for wake, 46.9% for N1, 87.2% for N2, 79.8% for N3, and 91.4% for REM in the test set (Figure 6.3 C).

The accuracy of the single-channel approach was 86.3% ($\kappa = 0.82$), 83.4% ($\kappa = 0.78$), and 82.9% ($\kappa = 0.77$) in the training, validation, and test sets, respectively. In the test set, the accuracies were 89.8% for wake, 46.0% for N1, 86.5% for N2, 75.4% for N3, and 90.8% for REM (Figure 6.3 D).

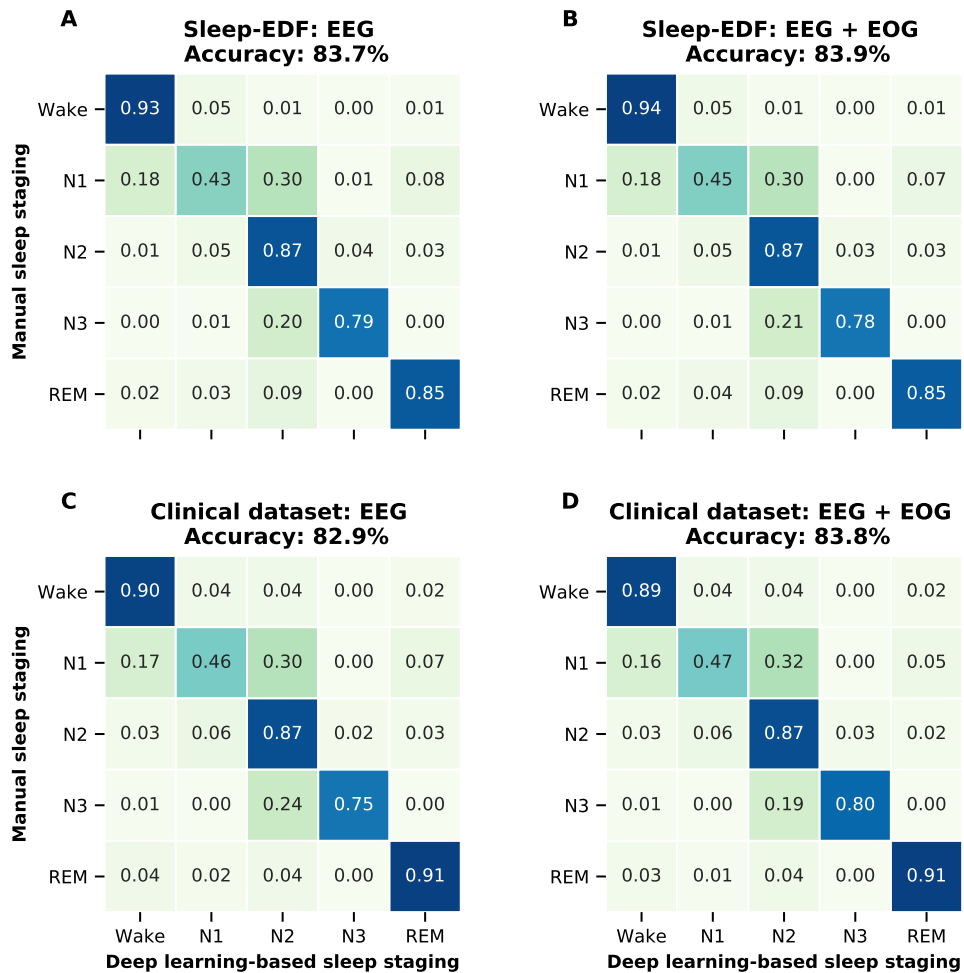


Figure 6.3: Normalized confusion matrices of the deep learning-based sleep staging with (A) single-channel input (EEG: Fpz-Cz) and (B) two-channel input (Fpz-Cz EEG and horizontal EOG) in the Sleep-EDF dataset as well as (C) single-channel input (EEG: F4-M1) and (D) two-channel input (EEG: F4-M1 and EOG: E1-M2) in the clinical dataset. Figure reconstructed based on study II.

Table 6.3: Comparison of the results obtained in study II with previous studies utilizing the Sleep-EDF dataset using cross-validation with an independent test set and having excluded the excess wake periods from the beginning and end of the recordings.

	Accuracy	κ	Recordings	Cross-validation
<i>Single-channel: Fpz-Cz</i>				
Study II	83.7%	0.77	153	10-fold
Phan et al. [132]	82.6%	0.76	153	10-fold
Mousavi et al. [121]	80.0%	0.73	153	10-fold
Mousavi et al. [121]	84.3%	0.79	39	20-fold
Supratak et al. [122]	82.0%	0.76	39	20-fold
Phan et al. [123]	81.9%	0.74	39	20-fold
Tsinalis et al. [124]	78.9%	-	39	20-fold
Tsinalis et al. [125]	74.8%	-	39	20-fold
<i>Two-channel: Fpz-Cz and EOG</i>				
Study II	83.9%	0.78	153	10-fold
Phan et al. [123]	82.3%	0.75	39	20-fold
Andreotti et al. [127]	76.8%	0.68	38	20-fold

6.2.3 Effect of OSA severity on sleep staging

When investigating the effect of OSA severity on the sleep staging performance via ten-fold cross-validation within each OSA severity category, the results revealed that the accuracy decreased with increasing OSA severity (Table 6.4). The accuracy and kappa values were highest for those individuals without OSA and lowest in severe OSA patients. Similarly, individual sleep stage accuracies were generally lowest in patients with severe OSA; however, the N1 accuracy in severe OSA patients was the highest of all the severity categories (Figure 6.4).

Table 6.4: The performance of the automatic sleep staging in patients without obstructive sleep apnea (OSA) ($n = 152$), with mild OSA ($n = 278$), with moderate OSA ($n = 208$), and with severe OSA ($n = 254$).

	Accuracy			Cohen's kappa (κ)		
	Training	Validation	Test	Training	Validation	Test
Non-OSA	89.4%	84.4%	84.5%	0.86	0.79	0.79
Mild OSA	87.7%	82.4%	82.8%	0.83	0.77	0.77
Moderate OSA	87.2%	83.0%	82.2%	0.83	0.77	0.76
Severe OSA	82.9%	76.7%	76.5%	0.77	0.68	0.68

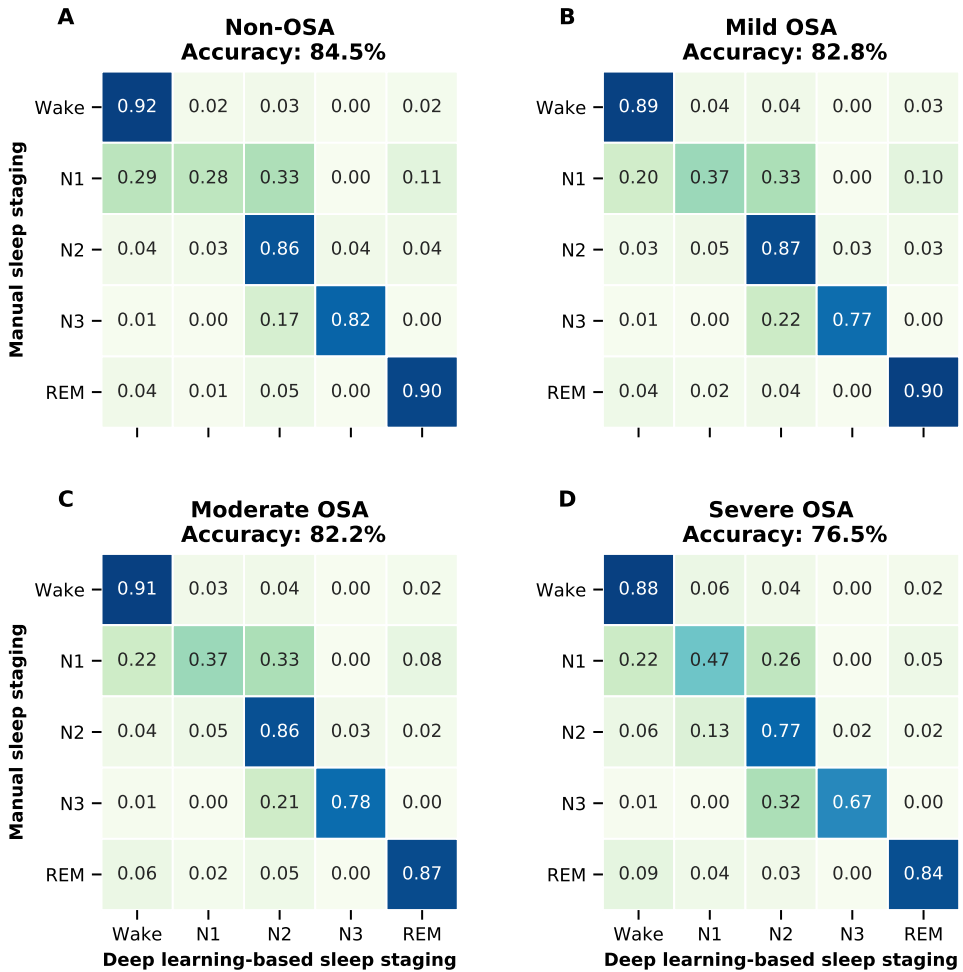


Figure 6.4: Normalized confusion matrices of the automatic sleep staging with a single EEG channel (F4-M1) in (A) non-OSA, (B) mild OSA, (C) moderate OSA, and (D) severe OSA patients. Figure reconstructed based on study II.

6.3 DEEP LEARNING-BASED AUTOMATIC SLEEP STAGING BASED ON PHOTOPLETHYSMOGRAM

In study III, sleep staging was conducted based solely on a photoplethysmogram (PPG) signal recorded with a finger pulse oximeter using three different classification systems: 3-stage classification (wake/NREM/REM), 4-stage classification (wake/light sleep/deep sleep/REM), and 5-stage classification (wake/N1/N2/N3/REM). Furthermore, total sleep time and sleep efficiency derived from the PPG-based sleep staging were compared to the manual scoring. Additionally, the AHI was calculated based on PPG sleep staging (PPG-AHI) by using the derived total sleep time and discarding respiratory events occurring during the epochs identified as wakefulness. This was then compared with the manual polysomnography-based AHI (PSG-AHI) and the corresponding to polygraphy-based AHI (PG-AHI) without information about the sleep stages.

6.3.1 Sleep staging accuracy

In the 3-stage classification (wake/NREM/REM), the PPG-based sleep staging achieved an epoch-by-epoch accuracy of 89.0% ($\kappa = 0.81$) in the training set, 79.5% ($\kappa = 0.63$) in the validation set, and 80.1% ($\kappa = 0.65$) in the independent test set. In the test set, wake was identified with 72.0% accuracy, NREM sleep with 87.1% accuracy, and REM sleep with 69.5% accuracy (Figure 6.5 A).

The 4-stage classification (wake/light sleep/deep sleep/REM) yielded an accuracy of 83.1% ($\kappa = 0.75$) in the training set, 67.1% ($\kappa = 0.51$) in the validation set, and 68.5% ($\kappa = 0.54$) in the independent test set. The model classified wake, light sleep, deep sleep, and REM sleep with accuracies of 72.8%, 71.5%, 52.0%, and 66.9%, respectively (Figure 6.5 B).

In the 5-stage classification (wake/N1/N2/N3/REM), the accuracies were 77.5% ($\kappa = 0.69$) in the training set, 62.3% ($\kappa = 0.48$) in the validation set, and 64.1% ($\kappa = 0.54$) in the test set. For the individual sleep stages, the accuracies were 77.6% for wake, 12.5% for N1, 67.3% for N2, 53.7% for N3, and 68.8% for REM sleep (Figure 6.5 C). Figure 6.6 illustrates an example of the 30-second PPG signal epochs during correctly classified sleep stages.

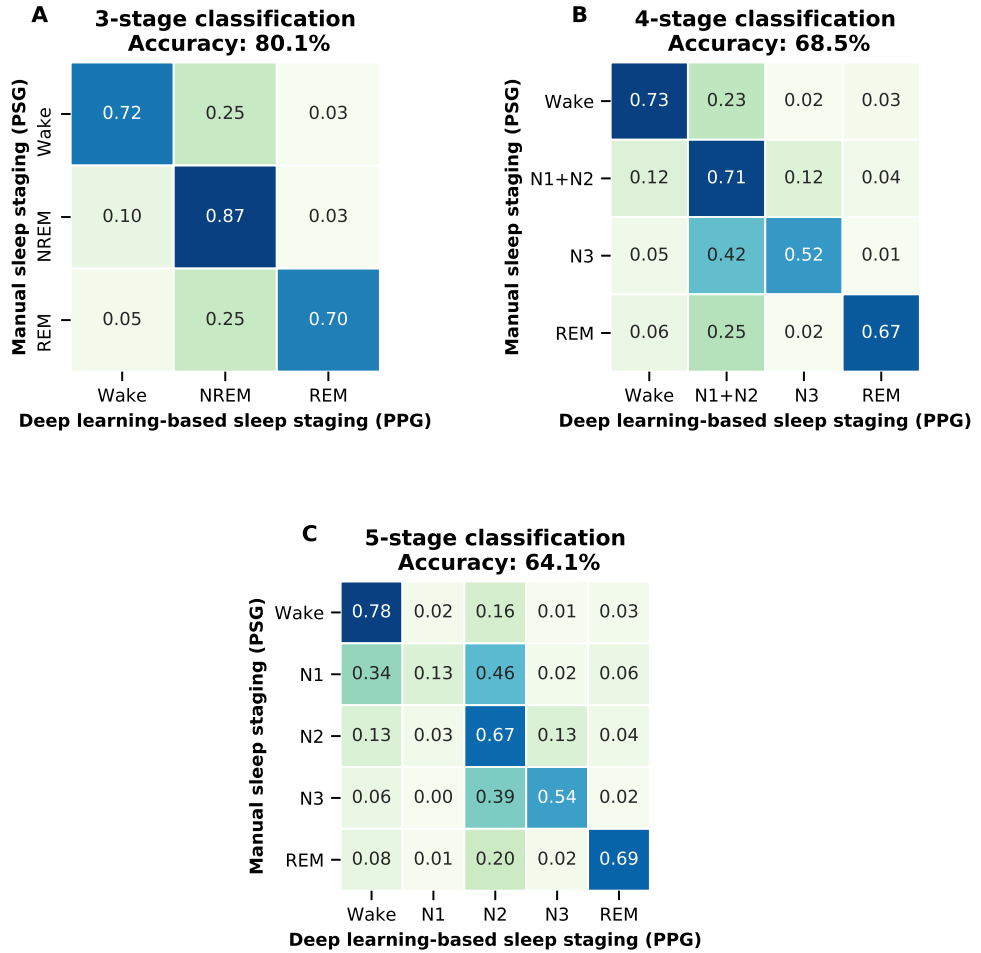


Figure 6.5: Normalized confusion matrix of the PPG-based sleep staging in (A) 3-stage classification (wake/NREM/REM), (B) 4-stage classification (wake/light sleep/deep sleep/REM), and (C) 5-stage classification (wake/N1/N2/N3/REM). Figure reconstructed based on study III.

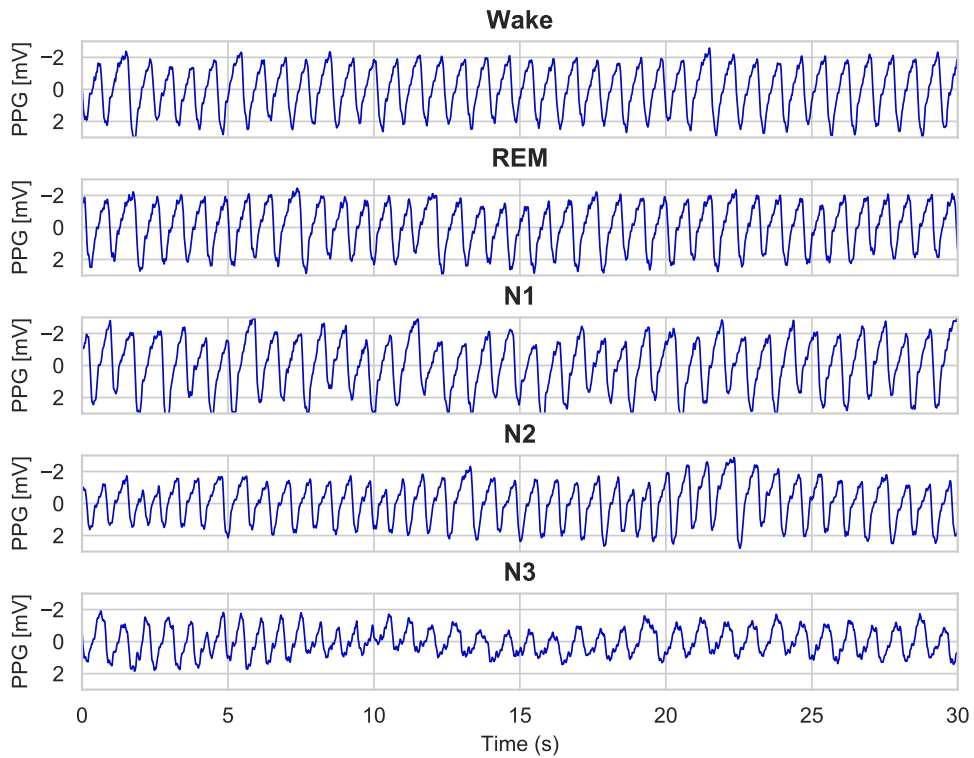


Figure 6.6: Examples of the PPG signals during correctly identified sleep stages. Stable PPG signal with a relatively constant frequency and amplitude can be observed during wakefulness. In contrast, the frequency decreases and an irregular variation in the signal amplitude becomes visible during N1 sleep. The amplitude further decreases and low-frequency oscillations in the signal begin to be evident when proceeding to N2 and N3 sleep. When comparing to wake, REM sleep is rather similar but with slightly higher variation in the amplitude. Figure reconstructed based on study III.

6.3.2 Derived clinical parameters

The smallest difference in derived total sleep time (TST) and sleep efficiency (SE) between manual sleep staging and PPG-based sleep staging was obtained with the 5-stage classification (Table 6.5). The Bland-Altman plots for the TST and SE are presented in Figure 6.7. Furthermore, the PG-AHI exhibited the largest difference when compared to the diagnostic PSG-AHI (Table 6.6), whereas the PPG-AHI was considerably closer to the diagnostic AHI, with the smallest difference being obtained with the 3-stage classification.

Table 6.5: The total sleep time and sleep efficiency derived from the manual sleep staging and the PPG-based sleep staging with 3-, 4-, and 5-stage classification. The mean (standard deviation) of the parameters and the mean difference to the manual sleep staging is presented.

	Total sleep time (min)			Sleep efficiency (%)		
	Mean	Mean difference	<i>p</i>	Mean	Mean difference	<i>p</i>
Manual	298.4 (79.8)	-	-	68.4 (16.9)	-	-
PPG: 3-stage	310.6 (76.1)	12.2 (52.9)	0.03	71.2 (15.7)	2.8 (11.3)	0.03
PPG: 4-stage	307.2 (78.7)	8.8 (55.5)	0.06	70.4 (16.5)	2.0 (12.0)	0.06
PPG: 5-stage	290.9 (81.9)	-7.5 (55.2)	0.24	66.6 (17.2)	-1.9 (12.2)	0.23

Table 6.6: The apnea-hypopnea index (AHI) derived from the manual scoring of polysomnography, the AHI simulated to present polygraphy, and AHI derived from the PPG-based sleep staging with 3-,4-, and 5-stage classifications. The mean (standard deviation) of the AHI values and the mean difference to the polysomnography-based AHI are presented.

	Apnea-hypopnea index (h^{-1})		
	Mean	Mean difference	<i>p</i>
Polysomnography	24.2 (24.3)	-	-
Polygraphy	18.8 (17.5)	-5.3 (12.4)	<0.001
PPG: 3-stage	23.3 (22.5)	-0.9 (9.0)	0.005
PPG: 4-stage	23.1 (22.1)	-1.1 (8.5)	0.002
PPG: 5-stage	22.6 (22.0)	-1.6 (8.5)	0.001

The AHI from PPG was calculated based on the total sleep time derived from the automatic sleep staging and by discarding respiratory events occurring during wake epochs. The polygraphy AHI was calculated by including all the respiratory events regardless of the prevalent sleep stage and dividing by the total recording time.

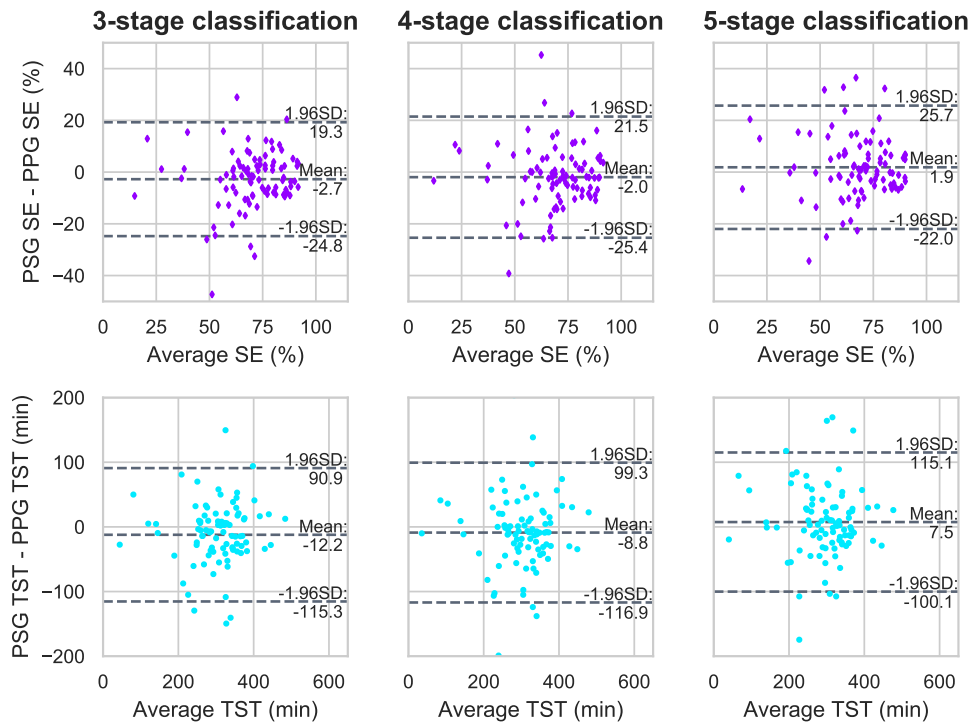


Figure 6.7: Bland-Altman plots for the total sleep time (TST) and sleep efficiency (SE) from the photoplethysmogram-based automatic sleep staging with 3-, 4-, and 5-stage classifications. The plots represent the average parameter values and the difference between the values based on the manual and the automatic sleep staging. Figure reconstructed based on study III.

6.4 DETAILED ANALYSIS OF SLEEP ARCHITECTURE WITH DEEP LEARNING

In study IV, the deep learning-based automatic sleep staging developed in study II was applied to identify sleep stages with varying degrees of overlap between the 30-second epochs to enable the deep learning-based method to assess the sleep architecture in more detail and to better detect the transitions between sleep stages. The sleep architecture and sleep fragmentation of individuals in each OSA severity group determined with the deep learning-based sleep staging were compared using the traditional non-overlapping 30-second epochs and with 15-, 5-, or 1-second epoch-to-epoch durations.

6.4.1 Sleep stage percentages and sleep parameters

In the severe OSA group, the amount of wake and N1 increased with shorter epoch-to-epoch durations revealing the disrupted sleep architecture characterized by numerous awakenings during the night (Table 6.7). Simultaneously, the amount of deep sleep decreased with shorter epoch-to-epoch durations. Conversely, in the groups with non-OSA, mild OSA, and moderate OSA, the percentage of wake and N1 decreased while the amount of N3 increased with shorter epoch-to-epoch durations. This OSA-related sleep fragmentation could not be captured when identifying sleep stages automatically with non-overlapping epochs. Thus, the more detailed assessment of sleep architecture with shorter epoch-to-epoch durations revealed major differences in the sleep architecture especially between severe OSA and non-OSA groups. An example of sleep stages identified with the traditional approach along with varying the epoch-to-epoch durations is illustrated in Figure 6.8.

Similar differences were found when assessing sleep parameters (total sleep time, TST; sleep efficiency, SE; wake after sleep onset, WASO) derived from the deep learning-based sleep staging using the traditional approach and with shorter epoch-to-epoch durations (Table 6.8). In the severe OSA group, the shorter epoch-to-epoch durations decreased the values of the TST and SE while increasing the WASO. In the non-OSA group, the effect was in the opposite pattern: TST and SE increased and WASO decreased with shorter epoch-to-epoch durations.

Table 6.7: The sleep stage percentages with varying epoch-to-epoch durations in obstructive sleep apnea (OSA) severity groups.

Epoch-to-epoch duration	Wake (% of recording)	N1 (% of sleep)	N2 (% of sleep)	N3 (% of sleep)	REM (% of sleep)
<i>Non-OSA</i>					
Original: 30 s	33.3	11.6	50.1	19.3	19.0
15 s	31.5	9.6	52.6	19.9	17.8
5 s	27.9	6.1	54.5	22.1	17.3
1 s	26.9	4.9	55.0	22.4	17.7
<i>Mild OSA</i>					
Original: 30 s	32.4	11.9	52.0	16.8	19.3
15 s	32.5	9.5	54.3	18.4	17.8
5 s	29.6	6.8	55.8	20.4	17.0
1 s	27.3	5.7	57.1	20.6	16.6
<i>Moderate OSA</i>					
Original: 30 s	31.5	12.1	52.8	16.8	18.4
15 s	31.7	9.8	54.1	18.4	17.7
5 s	31.9	8.0	56.5	18.7	16.7
1 s	30.2	7.3	57.9	18.4	16.4
<i>Severe OSA</i>					
Original: 30 s	32.0	12.2	52.5	17.1	18.1
15 s	33.5	12.1	54.7	15.3	17.8
5 s	36.8	14.8	56.2	13.3	15.7
1 s	35.6	15.5	56.0	13.6	14.8

Table 6.8: Total sleep time, wake after sleep onset, and sleep efficiency with varying epoch-to-epoch durations in obstructive sleep apnea (OSA) severity groups.

Epoch-to-epoch duration	Total sleep time (min)	Wake after sleep onset (min)	Sleep efficiency (%)
<i>Non-OSA</i>			
Original: 30 s	296.3 (78.3)	138.7 (66.0)	66.6 (15.5)
15 s	304.1 (77.4)	131.7 (67.6)	68.4 (15.6)
5 s	319.9 (73.2) [†]	119.1 (63.7) [†]	72.0 (14.8) [†]
1 s	324.4 (78.7) [†]	116.2 (68.3) [†]	72.9 (15.9) [†]
<i>Mild OSA</i>			
Original: 30 s	300.3 (79.0)	135.5 (74.0)	67.7 (16.7)
15 s	300.0 (83.4)	137.7 (74.0)	67.4 (17.0)
5 s	312.6 (75.5)	127.8 (70.0)	70.4 (15.3)
1 s	323.1 (77.1) [†]	117.5 (71.0) [†]	72.8 (15.6) [†]
<i>Moderate OSA</i>			
Original: 30 s	305.1 (76.4)	133.5 (71.2)	68.6 (15.6)
15 s	304.5 (70.6)	135.5 (71.3)	68.6 (15.0)
5 s	303.7 (73.5)*	137.2 (64.4)*	68.1 (14.3)*
1 s	311.2 (77.1)*	129.9 (69.1)*	69.7 (15.4)*
<i>Severe OSA</i>			
Original: 30 s	299.3 (73.1)	134.4 (69.7)	68.1 (15.1)
15 s	292.8 (77.3)	142.7 (70.3)	66.4 (15.7)
5 s	278.4 (75.8)* [†]	158.3 (74.2)* [†]	63.4 (16.3)* [†]
1 s	283.7 (82.2)*	152.6 (80.3)* [†]	64.6 (17.9)*

A statistically significant difference ($p < 0.05$) compared to the traditional sleep staging with non-overlapping epochs is denoted with a dagger (†) and between OSA severity groups when compared to the group without OSA is denoted with an asterisk (*).

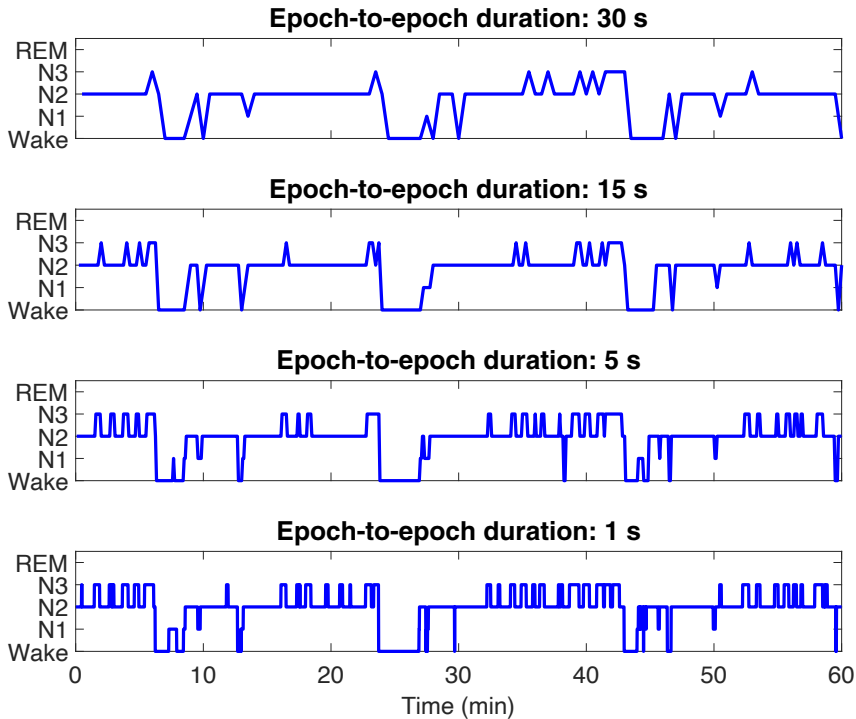


Figure 6.8: Examples of sleep stages scored with the traditional, non-overlapping epochs and scored with decreasing epoch-to-epoch duration between epochs. Figure reconstructed based on study IV.

6.4.2 Assessing sleep fragmentation via survival analysis

No significant differences in sleep fragmentation between the non-OSA and OSA severity groups were visible when the deep learning-based sleep staging was conducted with traditional non-overlapping epochs. The hazard ratios illustrating the risk of fragmented sleep (i.e. a short mean duration of continuous sleep periods) were close to one and were not statistically significant for any of the OSA severity groups when using non-overlapping epochs. Furthermore, no notable differences in the Kaplan-Meier survival curves (Figure 6.9) were observed. When the sleep fragmentation was evaluated based on the more detailed sleep staging with shorter epoch-to-epoch durations, differences emerged between the non-OSA and the OSA severity groups. The Kaplan-Meier survival curves (Figure 6.9) revealed differences in the sleep fragmentation between the groups; the mean duration of sleep periods decreased with increasing OSA severity. Similarly, the hazard ratios for fragmented sleep increased with decreasing epoch-to-epoch duration (Table 6.9). With the 1-second epoch-to-epoch duration, the hazard ratios were 1.21 ($p = 0.06$), 1.67 ($p < 0.01$), and 3.90 ($p < 0.01$) in the mild, moderate, and severe OSA groups.

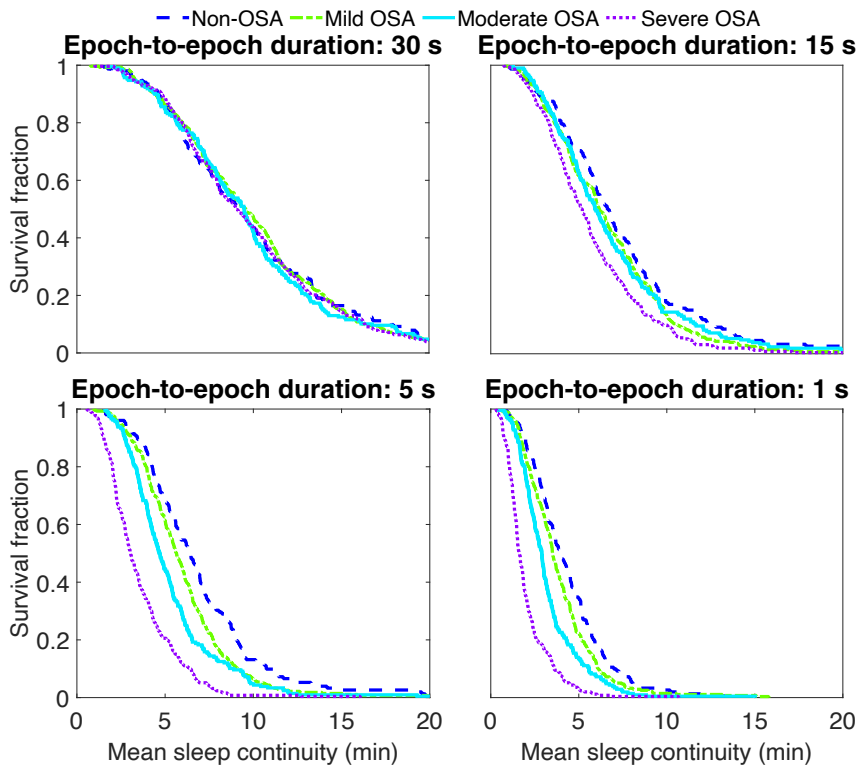


Figure 6.9: Kaplan-Meier survival curves representing the sleep continuity with different epoch-to-epoch durations. The curves illustrate the mean duration of continuous sleep in each obstructive sleep apnea group (OSA). Figure reconstructed based on study IV.

Table 6.9: The hazard ratios for fragmented sleep in obstructive sleep apnea (OSA) severity groups.

Epoch-to-epoch duration	Hazard ratio	95% confidence interval	<i>p</i>
<i>Mild OSA</i>			
Original: 30 s	1.00	0.82-1.21	0.97
15 s	1.20	0.98-1.46	0.08
5 s	1.31	1.07-1.59	0.01
1 s	1.21	0.99-1.48	0.06
<i>Moderate OSA</i>			
Original: 30 s	1.05	0.85-1.29	0.66
15 s	1.14	0.93-1.41	0.21
5 s	1.64	1.33-2.03	<0.01
1 s	1.67	1.35-2.06	<0.01
<i>Severe OSA</i>			
Original: 30 s	1.05	0.86-1.29	0.61
15 s	1.58	1.29-1.93	<0.01
5 s	3.54	2.88-4.36	<0.01
1 s	3.90	3.17-4.81	<0.01

The hazard ratios for fragmented sleep were calculated using the Cox proportional hazards model where the studied event was awakening from sleep and the time to event was the mean duration of sleep periods for each individual. Within each epoch-to-epoch duration, the sleep fragmentation is compared to the non-OSA group.

7 Discussion

Study I revealed that the AHI thresholds currently used in the severity classification of OSA are not optimal while the threshold combination of 3-9-24 h⁻¹ could improve the classification so that it would better reflect the OSA-related risk of all-cause mortality. Study II presented a deep learning-based sleep staging method with an accuracy comparable to manual sleep staging, even with only a single frontal EEG channel. Study III further demonstrated that sleep staging is possible based solely on a photoplethysmogram (PPG) signal measured with a simple finger pulse oximeter. Study IV further implemented the deep learning approach to achieve a more detailed analysis of sleep architecture. This revealed the highly fragmented sleep architecture of OSA patients which remains heavily overlooked with the traditional sleep staging.

7.1 OPTIMIZING THE SEVERITY ASSESSMENT OF OSA

In study I, the AHI-based severity classification of OSA was optimized with regards to the risk of all-cause mortality in a retrospective follow-up of a large population ($n = 1783$, mean follow-up time = 18.3 years) of patients having undergone an ambulatory polygraphy. The results indicated that the current threshold combination of 5-15-30 h⁻¹ is not optimal for assessing OSA severity. Instead, the AHI threshold combination of 3-9-24 h⁻¹ would better differentiate patients into OSA severity categories with regards to the risk of all-cause mortality. These optimized thresholds would ensure that an increase in the OSA severity can be correlated with an increase in the risk of all-cause mortality. However, the hazard ratios representing the risk of all-cause mortality varied greatly across the threshold simulations. This emphasizes that the AHI is not the sole factor determining the severity of OSA and thus more comprehensive measures of disease severity are needed.

The main determining factor for the magnitude of the hazard ratios for all the OSA severity categories was the threshold for dividing between the non-OSA and mild OSA categories. The results indicate that this threshold should be lowered from the current AHI value of 5 h⁻¹, as the hazard ratios increased when reducing this threshold. The results suggest that the patients with an AHI between 3 and 5 h⁻¹ also suffer from an elevated risk of OSA-related all-cause mortality. After lowering this threshold to 3 h⁻¹, the overall risk of all-cause mortality decreased in the non-OSA category which in turn increased the hazard ratios in the other OSA severity categories. However, if the threshold value was decreased below 3 h⁻¹, the hazard ratios of all severity categories relative to the non-OSA category began to decrease as patients with a low risk of OSA-related all-cause mortality were now classified as having mild OSA. Furthermore, the results suggest that the thresholds between mild and moderate, and between moderate and severe OSA should be lowered to 9 h⁻¹ and 24 h⁻¹, respectively. This would enable better differentiation of patients into OSA severity categories and achieve a more linear increase in the risk of all-cause mortality when progressing towards more severe OSA.

The utilization of the optimized OSA thresholds could significantly change the OSA severity assessment and treatment decisions. The thresholds used to classify OSA severity are currently applied to provide information about the disease to patients and to guide the treatment decision making. Rigorous thresholds and guidelines are required to persuade payers of subsidized therapy in insurance companies and healthcare systems. For example, in some healthcare systems, only patients with moderate-to-severe OSA can receive government- or health insurance-subsidized therapy. With the optimized threshold combination of 3-9-24 h⁻¹ the percentage of patients belonging to the moderate-to-severe OSA category increased from 27% to 39% in our studied population. Therefore, this classification could increase the number of patients receiving treatment thus elevating treatment-related healthcare costs. However, Ronald et al. [49] have previously shown that healthcare costs of OSA patients are higher than those of matched individuals without OSA, even before the OSA diagnosis. Furthermore, Albarrak et al. [50] demonstrated that these elevated costs of OSA patients become reduced to the same level as the healthy population after OSA treatment. Therefore, lowering the OSA severity thresholds could, in reality, reduce the long-term costs for healthcare systems by reducing OSA-related health consequences while improving the quality of life of OSA patients. However, further studies are warranted to thoroughly assess the healthcare costs after the utilization of the proposed lower severity thresholds.

The hazard ratios for the OSA severity categories were strongly dependent on the thresholds and varied significantly across different combinations. The most likely explanation behind this phenomenon is that the AHI does not fully capture the OSA severity and the risk of OSA-related health consequences. This proposition is also supported by previous studies; it has been shown that patients with similar AHI values can suffer from different degrees of physiological consequences caused by the respiratory events, for example, different depths and durations of the respiratory event-related oxygen desaturations [157]. Furthermore, patients with similar AHI values can have a different OSA phenotype and a differing risk of various severe health consequences [117]. Previous studies have also illustrated that it is the severity of desaturations that better reflects the OSA-related sleepiness and impaired vigilance rather than the AHI [6, 43]. Therefore, it might be beneficial to assess the severity of OSA in a more individualized manner by considering the physiological effects, risk factors, and health outcomes rather than simply the number of the respiratory events. For example, even though OSA patients suffer from highly fragmented sleep, as was reported in study IV, the extent of sleep fragmentation is not considered when assessing the severity of OSA. Furthermore, the severity of the desaturations is also completely overlooked when classifying patients into OSA severity categories. Therefore, the OSA severity may be better captured by supplementing the current AHI-based classification with an assessment of sleep fragmentation and sleep quality and with the severity of hypoxemia and other physiological effects of respiratory events. Therefore, further studies are warranted to improve the OSA severity classification as well as optimizing the AHI-based severity classification.

The main limitation of study I is that the thresholds were optimized only for ambulatory polygraphy scored using a 4% desaturation threshold for identifying hypopneas. As PSGs enable the identification of arousal-related hypopneas, the total number of respiratory events is higher as compared to polygraphy. This effect can also be observed from the results of study III, where the mean PSG-AHI was

24.2 h⁻¹ whereas the PG-AHI was 18.8 h⁻¹ (Table 6.6). Therefore, it could be argued that the optimal thresholds for PSG would be higher than the suggested 3-9-24 h⁻¹. Furthermore, scoring the hypopneas by using a 3% desaturation threshold affects the number of identified respiratory events and it has been shown that this causes the AHI to significantly increase [52, 53]. Thus, it is reasonable to assume that when using the 3% desaturation threshold, the optimal thresholds would be higher than with the suggested 3-9-24 h⁻¹. However, it has been suggested that the thresholds used for OSA severity classification should not remain the same for PG and PSG recordings nor with different desaturation thresholds [54, 158]. Therefore, the optimized thresholds of 3-9-24 h⁻¹ could only be considered as optimized for polygraphic studies and a 4% desaturation threshold. Further studies are warranted to perform a similar simulation-based approach to optimize the thresholds for PSGs, a 3% desaturation threshold and also for other outcomes in addition to the risk of all-cause mortality.

7.2 DEEP LEARNING-BASED SLEEP STAGING

In study II, the accuracy of the developed deep learning-based method for automatic sleep staging based on EEG and EOG compared favourably to previous automatic sleep staging approaches and to the inter-rater reliability between two expert manual scorers. The Cohen's kappa coefficient (κ) illustrating inter-rater reliability between two manual scorers is generally around 0.76 [159] but can be as low as 0.58 to 0.63 between sleep centers [160, 161]. Furthermore, in a previous study conducted in the Princess Alexandra Hospital's Sleep Center, where the clinical dataset used in this study was collected, the mean (standard deviation) kappa was 0.74 (0.02) [162]. Therefore, the two-channel and single-channel kappa values of 0.78 and 0.77, respectively, indicate a similar agreement as achieved with manual sleep staging conducted by two manual scorers.

Furthermore, the developed deep learning-based sleep staging method surpassed the accuracy of previous automatic approaches when tested using the Sleep-EDF dataset [143, 144] (Table 6.3). However, only Mousavi et al. [121] and Phan et al. [132] have previously used the updated Sleep-EDF dataset comprising 153 recordings. Similarly as in study II, they both included only 30 minutes of wakefulness before and after the sleep period and used ten-fold cross-validation. Mousavi et al. [121] achieved an accuracy of 80.03% ($\kappa = 0.73$) with a single EEG channel and Phan et al. reported an accuracy of 82.6% ($\kappa = 0.76$). The accuracy achieved in study II was 83.7% ($\kappa = 0.77$). Other studies utilized the older version of the Sleep-EDF dataset with only 39 recordings [122–125, 127] which hinders any direct comparison with these studies. However, Mousavi et al. [121] compared the accuracy of their method between the updated and the original Sleep-EDF dataset and demonstrated that the accuracy was significantly higher with the smaller dataset (84.26% accuracy vs. 80.03% accuracy). Furthermore, it is noteworthy that the results are only comparable to studies that have removed long wakefulness periods before and after sleep and have used an independent test set during evaluation. Some previous studies have included the easily identifiable excessive wake periods and did not have an independent test set, and have thus obtained overly optimistic results distorted by overfitting. These methodologically weak studies were excluded from the comparison.

The PPG-based automatic method published in study III differs from earlier approaches to identify sleep stages without relying on EEG. Previous approaches have generally relied on features of the ECG-derived heart rate variability (HRV) [134] usually accompanied by a recording of overnight movement [136] and respiratory effort [135]. In addition to ECG, the HRV features can be estimated to some extent from the PPG signal, and therefore efforts have been made to exploit the PPG-estimated HRV features for identifying sleep stages [138–142]. Aside from relying on estimated features, these approaches have relied on a simultaneous actigraphy recording. However, changes in the PPG signal have been previously linked to sympathetic activation [87, 88], EEG power density, and cortical activity during sleep [87]. This supports the exploitation of the full PPG signals without manually derived features in an end-to-end approach for sleep staging, as conducted in study III.

The accuracy of the PPG-based sleep staging method developed in study III compared favourably with previous studies attempting sleep staging from PPG, even though these studies have examined only a relatively small number of healthy individuals ($n = 10\text{--}152$) [138–142]. In these previous studies, a 2-stage sleep-wake classification has been conducted with 72%–77% accuracy [140–142], the 3-stage classification (wake/NREM/REM) with an accuracy of 73% ($\kappa = 0.46$) [138], and the 4-stage classification (wake/light sleep/deep sleep/REM) with an accuracy of 59%–69% ($\kappa = 0.42\text{--}0.52$) [138, 139]. In comparison, the PPG-based approach developed in study III achieved an accuracy of 80% ($\kappa = 0.65$) in the 3-stage classification, and 69% ($\kappa=0.54$) in the 4-stage classification. Furthermore, the results compare favourably with the ECG-based approaches, which have achieved accuracies of 80%–82% ($\kappa = 0.56\text{--}0.63$) in the 3-stage, and 69%–75% ($\kappa = 0.49\text{--}0.54$) in the 4-stage classification [134, 135]. Therefore, the results obtained in study III illustrate that the PPG-based sleep staging can be conducted with good accuracy by using the complete PPG signals without relying on a simultaneous actigraphy recording, and furthermore that the method is applicable in OSA patients.

Sleep staging is generally problematic in individuals affected by OSA. The reliability of manual sleep staging based on PSGs is lower in OSA populations [163, 164]. This has also hindered previous automatic EEG-based approaches [128] as deep learning-based scoring learns from the manual analysis. The lower reliability is most likely due to the fragmented sleep structure and the increased amount of N1 sleep which is characteristic for OSA patients. The fragmented sleep structure and increased amount of N1 sleep can be further observed from Table 5.3 and the results of study IV (Table 6.9 and Figure 6.9). In study II, it was observed that the accuracy of the sleep staging decreased when progressing towards more severe OSA (Table 6.4 and Figure 6.4). For example, the accuracy in the non-OSA group was 84.5% declining to 76.5% in severe OSA patients. In the severe OSA group, the amount of N1 sleep was 15% of the recording, while in the individuals without OSA, the percentage of N1 was only 6%. The accuracy in identifying N1 was always the lowest of all sleep stages (Figure 6.4). The consistent decrease in epoch-by-epoch accuracy with increasing OSA severity is, therefore, most likely due to the increasing amount of N1 and the increased number of transitions between sleep stages during the night. The results of study IV also support this proposal, since it was observed that the sleep architecture of severe OSA patients was highly fragmented.

A more accurate automatic sleep staging could have significant benefits in a clinical setting. The main advantage of automatic sleep staging would be the capability to consistently identify the sleep stages. Furthermore, the reliability

was already on par with a manual sleep staging in study **II**, and the automatic approach was able to conduct the sleep staging for a single night in a matter of seconds. Currently, the manual identification of sleep stages displays poor inter-rater reliability, especially between several individual sleep centers [159–161, 163–165]. The automatic method always conducts the sleep staging consistently without being affected by human factors, such as vigilance level, scoring environment, or human error. Furthermore, the current clinical practice of manual sleep staging poses a significant workload on trained professionals forcing them to conduct a highly repetitive and tedious task. Therefore, the automatic sleep staging could alleviate the clinical burden and free the valuable time of healthcare professionals for more meaningful tasks such as interpreting the results and conveying the information to patients. Moreover, future directions should include automatic scoring of all the identified events (e.g. respiratory disruptions, leg movements) incorporated into a single entity. Ultimately, these could make it possible to provide more individualized treatment planning.

Furthermore, the methods developed in studies **II** and **III** together provide opportunities for a more comprehensive utilization of ambulatory polygraphic studies. Various ambulatory EEG acquisition systems have already been developed [93–95, 166] and have a considerable potential to be implemented together with the automatic sleep staging outside sleep laboratories. Furthermore, the PPG-based sleep staging approach developed in study **III** would not require any modifications to the currently conducted polygraphic studies. Generally, a PG signal is already recorded in most polygraphic studies with a finger pulse oximeter and the deep learning-based approach developed in study **III** could be easily integrated into these recordings. Therefore, the PPG-based sleep staging could significantly enhance the polygraphic recordings and increase their diagnostic yield. The accuracy of the PPG-based sleep staging was lower than a staging based on a single-channel EEG (64.1% vs 82.9%). Nevertheless, the mean difference in total sleep time was only 7.5 min when compared to manual sleep staging despite the presence of some outliers (Figure 6.7) that led to a relatively large standard deviation of 55.2 min. Furthermore, the PPG-AHI only differed from the PSG-AHI on average by -0.9 h^{-1} . Therefore, depending on the required level of accuracy, either the single-channel EEG- or PPG-based sleep staging approach could be an extremely useful way to gain insights into the sleep architecture and sleep quality as well as helping in the diagnosis of other sleep disorders based on polygraphy, for example, REM-related OSA. Furthermore, the PPG-based sleep staging would provide an inexpensive way to undertake a long-term monitoring of sleep and could be used to supplement the current actigraphy-based methods. However, both the EEG- and PPG-based approaches will require further studies to validate their performance when applied with ambulatory methods.

Besides the applications in a clinical setting, the developed sleep staging methods have the potential to be incorporated into various consumer health technology devices. Currently, a reflective PPG measurement is already included in consumer-grade wearable self-tracking devices. Some of these already claim to measure sleep duration and sleep quality but they lack clinical validation and the implemented algorithms are not in the public domain [167–170]. Conversely, the PPG-based sleep staging in study **III** already provided highly promising results in a clinical population of patients in whom there was a suspicion of OSA. This reveals the potential of PPG-based sleep staging in individuals affected by sleep disorders and could enable a simple solution for monitoring sleep quality and identifying

sleep disorders, even with consumer-grade devices. However, as the reflective PPG measurement differs fundamentally from a transmissive measurement and causes additional challenges, further studies will be necessary to validate the performance of the deep learning-based method published in study III with data collected using reflective sensors integrated into consumer-grade wearable devices. Similarly, even the EEG-based sleep staging could become feasible in consumer-grade health technology devices measuring a limited number of EEG channels. In these devices, the integration of the deep learning-based sleep staging could enable the long-term monitoring of sleep and with further development, identify abnormal sleep architecture. However, these possibilities will require further validation with data collected using consumer-grade technologies.

The most significant limitation of studies II and III was the low agreement with manual sleep staging in detecting N1 sleep. With the EEG-based approach, the accuracy in detecting N1 varied between 28% and 47%, while the accuracy in the PPG-based approach was 13%. However, the agreement between manual scorers is similarly at its lowest when identifying the N1 sleep stage [163, 165]: the kappa value representing the N1 agreement of manual scorers is only between 0.19 and 0.46 [159–161]. Therefore, the low N1 agreement may be caused by the scoring rules for N1 and the extensive disagreement between manual scorers. Furthermore, the lower N1 accuracy with the PPG-based sleep staging when compared to the EEG-based approach may be due to insufficiently small differences in the PPG signal between the N1 and N2 stages, as the N1 was usually misidentified as N2. These results raise the question of whether the differentiation of N1 from N2 would actually be required for all applications. Aside from this, the current sleep staging practice suffers from arbitrary rules which do not have extensive foundation on physiological factors [13, 146, 171]. Furthermore, the use of non-overlapping 30-second epochs significantly overlooks many transitions between stages, which was also observed in study IV. Thus, the agreement with the manual sleep staging of PSG does not necessarily fully reflect the true accuracy of the automatic sleep staging method; thus, further studies are warranted to determine how the EEG- and PPG-based sleep staging approaches can capture the physiological changes occurring during the night and how they reflect the outcomes of sleep quality, for example, the presence of daytime sleepiness and deteriorated vigilance.

7.3 DETAILED ANALYSIS OF SLEEP ARCHITECTURE

In study IV, the deep learning-based sleep staging approach developed in study II was implemented to analyse sleep architecture in a more detailed manner. The automatic sleep staging from study II based on EEG and EOG channels was applied to identify sleep stages by taking new 30-second epochs with varying epoch-to-epoch durations and allowing for an overlap between consecutive epochs. The main result from study IV was that the more detailed sleep staging with shorter epoch-to-epoch duration revealed the highly disrupted sleep architecture of OSA patients which had been severely underestimated when the sleep staging was conducted with traditional non-overlapping epochs.

Shorter epoch-to-epoch durations revealed larger differences in the sleep stage percentages and sleep parameters (total sleep time, TST; sleep efficiency, SE; and wake after sleep onset, WASO) between the population without OSA and those individuals with severe OSA in comparison with the traditional non-overlapping

epochs. In the non-OSA group, the amount of wakefulness and REM decreased with shorter epoch-to-epoch durations, while the amount of N2 and N3 increased. Furthermore, a similar effect was observed in the mild and moderate OSA groups. In contrast, the N1 and wakefulness increased while N3 decreased in the severe OSA group with shorter epoch-to-epoch durations. Similarly, TST and SE decreased and WASO increased in the severe OSA group with shorter epoch-to-epoch durations with an opposite effect being evident in the population without OSA. These results illustrate that severe OSA patients suffer from more disrupted sleep than can be estimated when the sleep staging is conducted with traditional non-overlapping epochs.

Sleep staging with shorter epoch-to-epoch durations revealed differences in the sleep fragmentation between all of the OSA severity groups and the non-OSA group. When the deep learning model identified sleep stages with traditional non-overlapping epochs, neither the hazard ratios representing the risk of fragmented sleep (Table 6.9) nor the Kaplan-Meier survival curves (Figure 6.9) revealed any differences between the OSA severity groups. However, by applying overlapping epochs with shorter epoch-to-epoch durations, significant differences began to emerge; for example, with 1-second epoch-to-epoch duration, the hazard ratios representing sleep fragmentation were 1.21 ($p = 0.06$) for mild, 1.67 ($p < 0.01$) for moderate, and 3.90 ($p < 0.01$) for severe OSA. Similarly, the Kaplan-Meier curves revealed differences between all of the severity groups. It is noteworthy that the mean duration of continuous sleep periods within all severity groups and non-OSA group decreased with shorter epoch-to-epoch durations. However, this is an expected result as the sleep staging with shorter epoch-to-epoch durations provides a means to more comprehensively assess the sleep architecture while capturing more of the wake transitions during the night.

The more detailed assessment of sleep architecture revealed that the deep learning-based sleep staging with the traditional non-overlapping epochs underestimates the sleep fragmentation of severe OSA patients; however, this phenomenon was only seen to some extent in mild and moderate OSA patients. While the shorter epoch-to-epoch durations increased the differences in sleep stage percentages and sleep parameters between severe OSA and non-OSA groups, only small differences emerged between the mild or moderate OSA and the non-OSA groups. Overall, shortening of the epoch-to-epoch durations in the mild and moderate OSA groups produced similar effects than in the non-OSA group; for example, the percentage of N1 and wake decreased while that of N3 increased. However, differences between mild and moderate OSA and non-OSA groups emerged in the survival analysis-based assessment of sleep continuity, revealing that mild-to-moderate OSA causes sleep fragmentation but to a smaller extent than severe OSA. A similar phenomenon was observed by Norman et al. [42] who detected no significant differences between normal and mild OSA groups was observed in traditional sleep parameters whereas an assessment of sleep continuity via survival analysis was able to distinguish differences even between these two groups. One explanation for the small differences in sleep stage percentages and sleep parameters (TST, SE, WASO) between the mild or moderate OSA and the non-OSA groups could be the highly artificial classification into the severity groups as was observed in study I. As the more detailed sleep staging revealed differences in the sleep fragmentation between the OSA severity groups, this could provide a feasible way to assess OSA severity alongside the more conventional respiratory and hypoxemia-related methods. Thus, further studies incorporating these aspects

in OSA severity assessment are warranted.

Overall, the results of study **IV** illustrated that while automatic sleep staging with traditional non-overlapping epochs might be sufficient in healthy populations, a more detailed assessment is required when investigating individuals affected by sleep disorders or other disorders potentially affecting sleep. This is supported by the fact that the whole sleep staging process was originally designed based on a healthy population [12, 13]. The approach developed in study **IV** would provide a novel way to achieve a more realistic representation of the disrupted sleep architecture of individuals with sleep disorders, which appear to be seriously overlooked in the traditional sleep staging. Ultimately, this might result in a more informed, detailed, and individualized diagnosis of sleep disorders, which could enhance the severity classification of OSA, and even guide treatment planning. Additionally, an interesting topic for further studies would be determining the connection between the sleep architecture assessed via the more detailed approach and perceived sleep quality, daytime sleepiness and vigilance, therapeutic outcomes, and the risk of comorbidities.

Study **IV** was based on identifying sleep stages in overlapping 30-second epochs which is simultaneously the strength and limitation of the study. The sleep architecture determined with shorter epoch-to-epoch durations is interpretable similarly as the traditional sleep staging. The detailed approach enables the deep learning-based method to detect sleep stage transitions with a better temporal resolution, a factor which could be crucial in the diagnosis of sleep disorders. However, the study was still based on identifying discrete sleep stages for 30-second epochs and does not provide a continuous scale representing the depth of sleep. This has been previously attempted based on the frequency content of EEG [146, 171]. However, in these studies sleep depth was represented in an arbitrarily chosen continuous scale which cannot be as easily interpreted. Furthermore, in study **IV**, arousals from sleep were not considered in order to avoid relying on manual scoring which usually suffers from a low arousal scoring reliability [172]. Therefore, future studies are warranted to incorporate and compare the developed method alongside a continuous scale of sleep depth while also identifying and assessing arousals from sleep. Finally, the sleep staging was conducted with the deep learning-based approach developed in study **II**. The manual analysis was discarded to enable a reliable comparison between the different epoch-to-epoch durations. Therefore, the results can only be generalized to the deep learning-based sleep staging. No conclusions can be made about whether a manual analysis would behave similarly if sleep staging were to be conducted with shorter epoch-to-epoch durations instead of the traditional approach. However, manual analysis with overlapping epochs could be biased and impractical as the number of epochs that would require scoring would increase substantially as the changes between consecutive epochs became smaller.

8 Conclusions

The results included in this thesis demonstrated that the diagnostics of sleep disorders could be made more efficient and accurate with the application of deep learning. The assessment of obstructive sleep apnea severity can be optimized by simulating various threshold combinations and defining the severity of OSA based on the risk of severe health outcomes. Sleep staging, which is the cornerstone of sleep medicine, can be automatized with deep learning-based methods; these are able to reach clinical accuracy while remaining perfectly reproducible. With deep learning, sleep staging is possible with lighter measurement setups: a single EEG-channel or even a PPG measured with a simple finger pulse oximeter. Finally, the sleep architecture can be analysed in more detail by implementing deep learning-based sleep staging with shorter epoch-to-epoch durations. This can provide crucial insights into the sleep fragmentation occurring in individuals suffering from sleep disorders.

The main conclusions corresponding to the aims of the thesis are:

1. The current severity classification of obstructive sleep apnea based on the apnea-hypopnea index thresholds $5-15-30 \text{ h}^{-1}$ is not optimal in characterizing the risk of OSA-related all-cause mortality. These thresholds should be optimized for each measurement technique or the apnea-hypopnea index-based severity assessment needs to be supplemented with more comprehensive measures of the disease severity.
2. Sleep staging can be conducted automatically at an accuracy on par with a manual assessment conducted by expert somnologists by using deep learning methods. Accurate sleep staging is possible based on a single frontal EEG channel even in a clinical population of patients with suspected OSA. This could represent an easily applicable and cost-effective way of conducting sleep staging in clinical practice.
3. PPG measured with a simple finger pulse oximeter can be used for deep learning-based sleep staging. This enables a reasonably accurate determination of total sleep time and the sleep stages can be identified with a moderate agreement with the EEG-based sleep staging. As PPG is straightforward to record and has been already integrated into ambulatory recording setups, a deep learning-based sleep staging could achieve a cost-effective long-term assessment of sleep architecture based on home recordings which could significantly increase their diagnostic value.
4. Shorter epoch-to-epoch durations enabled the deep learning-based sleep staging to assess sleep architecture in more detail and thus reveal the highly fragmented sleep of patients suffering from severe OSA, which is easily overlooked with the traditional sleep staging. A more detailed assessment of sleep architecture would be paramount when assessing the sleep quality of individuals suffering from sleep disorders.

In summary, deep learning has immense potential to revolutionize the field of sleep medicine. With deep learning, the diagnostic processes can be made more efficient by reducing the number of signals that are required and by making the processes less reliant on the time-consuming and error-prone manual analysis. The diagnosis could become more consistent and informative by assessing the sleep architecture automatically and in more detail. Finally, the diagnostic yield of home-based measurements could be significantly increased and brought closer to the diagnostic accuracy of in-lab measurements.

BIBLIOGRAPHY

- [1] S. Diekelmann and J. Born, "The memory function of sleep," *Nature Reviews Neuroscience* **11**, 114–126 (2010).
- [2] N. E. Fultz, G. Bonmassar, K. Setsompop, R. A. Stickgold, B. R. Rosen, J. R. Polimeni, and L. D. Lewis, "Coupled electrophysiological, hemodynamic, and cerebrospinal fluid oscillations in human sleep," *Science* **366**, 628–631 (2019).
- [3] AASM, "Sleep-Related Breathing Disorders in Adults: Recommendations for Syndrome Definition and Measurement Techniques in Clinical Research," *Sleep* (1999).
- [4] A. V. Benjafield, N. T. Ayas, P. R. Eastwood, R. Heinzer, M. S. Ip, M. J. Morrell, C. M. Nunez, S. R. Patel, T. Penzel, J. L. D. Pépin, P. E. Peppard, S. Sinha, S. Tufik, K. Valentine, and A. Malhotra, "Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis," *The Lancet Respiratory Medicine* **7**, 687–698 (2019).
- [5] T. Penzel, J. W. Kantelhardt, C.-C. Lo, K. Voigt, and C. Vogelmeier, "Dynamics of heart rate and sleep stages in normals and patients with sleep apnea," *Neuropsychopharmacology* **28**, S48–S53 (2003).
- [6] S. Kainulainen, B. Duce, H. Korkalainen, A. Oksenberg, A. Leino, E. Arnardottir, A. Kulkas, S. Myllymaa, J. Töyräs, and T. Leppänen, "Severe desaturations increase psychomotor vigilance task-based median reaction time and number of lapses in obstructive sleep apnoea patients," *European Respiratory Journal* **55** (2020).
- [7] J. M. Marin, S. J. Carrizo, E. Vicente, and A. G. Agustí, "Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure: an observational study," *Lancet* **365**, 1046–1053 (2005).
- [8] V. K. Somers, D. P. White, R. Amin, W. T. Abraham, F. Costa, A. Culebras, S. Daniels, J. S. Floras, C. E. Hunt, L. J. Olson, T. G. Pickering, R. Russell, M. Woo, and T. Young, "Sleep apnea and cardiovascular disease: an American Heart Association/American College of Cardiology Foundation Scientific Statement from the American Heart Association Council for High Blood Pressure Research Professional Education Committee, Council on Clinical Cardiology, Stroke Council, and Council on Cardiovascular Nursing," *Journal of the American College of Cardiology* **52**, 686–717 (2008).
- [9] N. S. Marshall, K. K. H. Wong, S. R. J. Cullen, M. W. Knuiman, and R. R. Grunstein, "Sleep apnea and 20-year follow-up for all-cause mortality, stroke, and cancer incidence and mortality in the Busselton Health Study cohort," *Journal of Clinical Sleep Medicine* **10**, 355–362 (2014).

- [10] D. Hillman, S. Mitchell, J. Streatfeild, C. Burns, D. Bruck, and L. Pezzullo, "The economic cost of inadequate sleep," *Sleep* **41** (2018).
- [11] R. B. Berry, C. L. Albertario, S. M. Harding, R. M. Lloyd, D. T. Plante, S. F. Quan, M. M. Troester, and B. V. Vaughn, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications.*, Version 2.5 ed. (Darien, IL: American Academy of Sleep Medicine, 2018).
- [12] A. Rechtschaffen and A. Kales, *A manual of standardized terminology, techniques and scoring system of sleep stages in human subjects* (University of California, Brain Information Service/Brain Research Institute, Los Angeles, 1968).
- [13] S. L. Himanen and J. Hasan, "Limitations of Rechtschaffen and Kales," *Sleep Medicine Reviews* **4**, 149–167 (2000).
- [14] N. A. Collop, W. M. Anderson, B. Boehlecke, D. Claman, R. Goldberg, D. J. Gottlieb, D. Hudgel, M. Sateia, and R. Schwab, "Clinical guidelines for the use of unattended portable monitors in the diagnosis of obstructive sleep apnea in adult patients. Portable Monitoring Task Force of the American Academy of Sleep Medicine," *Journal of Clinical Sleep Medicine* **3**, 737–747 (2007).
- [15] R. B. Berry, R. Budhiraja, D. J. Gottlieb, D. Gozal, C. Iber, V. K. Kapur, C. L. Marcus, R. Mehra, S. Parthasarathy, S. F. Quan, S. Redline, K. P. Strohl, S. L. D. Ward, and M. M. Tangredi, "Rules for Scoring Respiratory Events in Sleep: Update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events," *Journal of Clinical Sleep Medicine* **8**, 597–619 (2012).
- [16] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala, "Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach," *Scientific Reports* **8**, 1–10 (2018).
- [17] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature* **542**, 115–118 (2017).
- [18] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* **521**, 436–444 (2015).
- [19] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning* (MIT press, 2016).
- [20] S. A. Mickelson, "Normal Sleep Physiology," Chap 1 in *Sleep Medicine*, K. L. Yaremchuk and P. A. Wardrop, eds. (Plural Publishing, 2011).
- [21] I. Feinberg and T. C. Floyd, "Systematic Trends Across the Night in Human Sleep Cycles," *Psychophysiology* **16**, 283–291 (1979).
- [22] H. R. Colten and B. M. Altevogt, *Sleep disorders and sleep deprivation : an unmet public health problem* (Institute of Medicine : National Academies Press, 2006).
- [23] P. Maquet, "The role of sleep in learning and memory," *Science* **294**, 1048–1052 (2001).
- [24] M. Dattilo, H. K. M. Antunes, A. Medeiros, M. M. Neto, H. S. d. Souza, S. Tufik, and M. De Mello, "Sleep and muscle recovery: endocrinological and molecular basis for a new and promising hypothesis," *Medical Hypotheses* **77**, 220–222 (2011).

- [25] M. W. Chee and Y. L. Chuah, "Functional neuroimaging and behavioral correlates of capacity decline in visual short-term memory after sleep deprivation," *Proceedings of the National Academy of Sciences* **104**, 9487–9492 (2007).
- [26] L. Besedovsky, T. Lange, and J. Born, "Sleep and immune function," *Pflügers Archiv-European Journal of Physiology* **463**, 121–137 (2012).
- [27] N. Goel, H. Rao, J. S. Durmer, and D. F. Dinges, "Neurocognitive consequences of sleep deprivation," in *Seminars in neurology*, Vol. 29 (Thieme Medical Publishers, 2009), pp. 320–339.
- [28] N. Tsuno, A. Besset, and K. Ritchie, "Sleep and depression," *The Journal of Clinical Psychiatry* **66**, 1254–1269 (2005).
- [29] E. Tobaldini, G. Costantino, M. Solbiati, C. Cogliati, T. Kara, L. Nobili, and N. Montano, "Sleep, sleep deprivation, autonomic nervous system and cardiovascular diseases," *Neuroscience & Biobehavioral Reviews* **74**, 321–329 (2017).
- [30] R. M. Benca, M. Okawa, M. Uchiyama, S. Ozaki, T. Nakajima, K. Shibui, and W. H. Obermeyer, "Sleep and mood disorders," *Sleep Medicine Reviews* **1**, 45–56 (1997).
- [31] P. M. Fuller, J. J. Gooley, and C. B. Saper, "Neurobiology of the sleep-wake cycle: sleep architecture, circadian regulation, and regulatory feedback," *Journal of Biological Rhythms* **21**, 482–493 (2006).
- [32] L. Marshall and J. Born, "The contribution of sleep to hippocampus-dependent memory consolidation," *Trends in Cognitive Sciences* **11**, 442–450 (2007).
- [33] R. W. McCarley, "Neurobiology of REM and NREM sleep," *Sleep Medicine* **8**, 302–330 (2007).
- [34] J. A. Hobson, "Sleep is of the brain, by the brain and for the brain," *Nature* **437**, 1254–1256 (2005).
- [35] B. Rasch, S. Gais, and J. Born, "Impaired off-line consolidation of motor memories after combined blockade of cholinergic receptors during REM sleep-rich sleep," *Neuropsychopharmacology* **34**, 1843–1853 (2009).
- [36] O. P. Hornung, F. Regen, H. Danker-Hopfe, M. Schredl, and I. Heuser, "The relationship between REM sleep and memory consolidation in old age and effects of cholinergic medication," *Biological Psychiatry* **61**, 750–757 (2007).
- [37] T. Penzel, J. W. Kantelhardt, L. Grote, J. H. Peter, and A. Bunde, "Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea," *IEEE Transactions on Biomedical Engineering* **50**, 1143–1151 (2003).
- [38] S. Elsenbruch, M. J. Harnish, and W. C. Orr, "Heart rate variability during waking and sleep in healthy males and females," *Sleep* **22**, 1067–1071 (1999).

- [39] V. K. Somers, M. E. Dyken, A. L. Mark, and F. M. Abboud, "Sympathetic-Nerve Activity during Sleep in Normal Subjects," *New England Journal of Medicine* **328**, 303–307 (1993).
- [40] I. Berlad, A. Shlitner, S. Ben-Haim, and P. Lavie, "Power spectrum analysis and heart rate variability in Stage 4 and REM sleep: evidence for state-specific changes in autonomic dominance," *Journal of Sleep Research* **2**, 88–90 (1993).
- [41] AASM, *International Classification of Sleep Disorders*, 3rd ed. (Darien, IL, 2014).
- [42] R. G. Norman, M. A. Scott, I. Ayappa, J. A. Walsleben, and D. M. Rapoport, "Sleep continuity measured by survival curve analysis," *Sleep* **29**, 1625–1631 (2006).
- [43] S. Kainulainen, J. Töyräs, A. Oksenberg, H. Korkalainen, S. Sefa, A. Kulkas, and T. Leppänen, "Severity of desaturations reflects OSA-related daytime sleepiness better than AHI," *Journal of Clinical Sleep Medicine* **15**, 1135–1142 (2019).
- [44] T. Young, P. E. Peppard, and D. J. Gottlieb, "Epidemiology of Obstructive Sleep Apnea: A Population Health Perspective," *American Journal of Respiratory and Critical Care Medicine* **165**, 1217–1239 (2002).
- [45] N. M. Punjabi, B. S. Caffo, J. L. Goodwin, D. J. Gottlieb, A. B. Newman, G. T. O'Connor, D. M. Rapoport, S. Redline, H. E. Resnick, J. A. Robbins, E. Shahar, M. L. Unruh, and J. M. Samet, "Sleep-disordered breathing and mortality: a prospective cohort study," *PLoS Medicine* **6**, e1000132 (2009).
- [46] T. Young, L. Finn, P. E. Peppard, M. Szklo-Coxe, D. Austin, F. J. Nieto, R. Stubbs, and K. M. Hla, "Sleep disordered breathing and mortality: eighteen-year follow-up of the Wisconsin sleep cohort," *Sleep* **31**, 1071 (2008).
- [47] G. J. Gibson, "Obstructive sleep apnoea syndrome: underestimated and undertreated," *British Medical Bulletin* **72**, 49–65 (2004).
- [48] F. . Sullivan, *Hidden health crisis costing America billions. Underdiagnosing and undertreating obstructive sleep apnea draining healthcare system* (Darien, IL: American Academy of Sleep Medicine, 2016).
- [49] J. Ronald, K. Delaive, L. Roos, J. Manfreda, A. Bahammam, and M. H. Kryger, "Health Care Utilization in the 10 Years Prior to Diagnosis in Obstructive Sleep Apnea Syndrome Patients," *Sleep* **22**, 225–229 (1999).
- [50] M. Albarrak, K. Banno, A. A. Sabbagh, K. Delaive, R. Walld, J. Manfreda, and M. H. Kryger, "Utilization of Healthcare Resources in Obstructive Sleep Apnea Syndrome: a 5-Year Follow-Up Study in Men Using CPAP," *Sleep* **28**, 1306–1311 (2005).
- [51] W. W. Flemons, N. J. Douglas, S. T. Kuna, D. O. Rodenstein, and J. Wheatley, "Access to Diagnosis and Treatment of Patients with Suspected Sleep Apnea," *American Journal of Respiratory and Critical Care Medicine* **169**, 668–672 (2004).
- [52] B. Duce, J. Milosavljevic, and C. Hukins, "The 2012 AASM Respiratory Event Criteria Increase the Incidence of Hypopneas in an Adult Sleep Center Population," *Journal of Clinical Sleep Medicine* **11**, 1425–1431 (2015).

- [53] S. Myllymaa, K. Myllymaa, S. Kupari, A. Kulkas, T. Leppänen, P. Tiihonen, E. Mervaala, J. Seppä, H. Tuomilehto, and J. Töyräs, "Effect of different oxygen desaturation threshold levels on hypopnea scoring and classification of severity of sleep apnea," *Sleep and Breathing* **19**, 947–954 (2015).
- [54] D. M. Rapoport, "POINT: Is the Apnea-Hypopnea Index the Best Way to Quantify the Severity of Sleep-Disordered Breathing? Yes," *Chest* **149**, 14–16 (2016).
- [55] T. Penzel, C. Schöbel, and I. Fietze, "Revise Respiratory Event Criteria or Revise Severity Thresholds for Sleep Apnea Definition?," *Journal of Clinical Sleep Medicine* **11**, 1357–1359 (2015).
- [56] L. J. Epstein, D. Kristo, J. S. P. J, N. Friedman, A. Malhotra, S. P. Patil, K. Ramar, R. Rogers, R. J. Schwab, E. M. Weaver, and M. D. Weinstein, "Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults," *Journal of Clinical Sleep Medicine* **5**, 263 (2009).
- [57] T. E. Weaver and R. R. Grunstein, "Adherence to continuous positive airway pressure therapy: The challenge to effective treatment," *Proceedings of the American Thoracic Society* **5**, 173–178 (2008).
- [58] A. Kulkas, T. Leppänen, J. Sahlman, P. Tiihonen, E. Mervaala, J. Kokkarinen, J. Randell, J. Seppä, H. Tuomilehto, and J. Töyräs, "Novel parameters reflect changes in morphology of respiratory events during weight loss," *Physiological measurement* **34**, 1013 (2013).
- [59] A. Oksenberg, D. Silverberg, D. Offenbach, and E. Arons, "Positional therapy for obstructive sleep apnea patients: a 6-month follow-up study," *The Laryngoscope* **116**, 1995–2000 (2006).
- [60] P. R. Eastwood, M. Barnes, J. H. Walsh, K. J. Maddison, G. Hee, A. R. Schwartz, P. L. Smith, A. Malhotra, R. D. McEvoy, J. R. Wheatley, et al., "Treating obstructive sleep apnea with hypoglossal nerve stimulation," *Sleep* **34**, 1479–1486 (2011).
- [61] D. J. Eckert, A. S. Jordan, P. Merchia, and A. Malhotra, "Central sleep apnea: pathophysiology and treatment," *Chest* **131**, 595–607 (2007).
- [62] R. N. Aurora, S. Chowdhuri, K. Ramar, S. R. Bista, K. R. Casey, C. I. Lamm, D. A. Kristo, J. M. Mallea, J. A. Rowley, R. S. Zak, et al., "The treatment of central sleep apnea syndromes in adults: practice parameters with an evidence-based literature review and meta-analyses," *Sleep* **35**, 17–40 (2012).
- [63] S. Schutte-Rodin, L. Broch, D. Buysse, C. Dorsey, and M. Sateia, "Clinical guideline for the evaluation and management of chronic insomnia in adults," *Journal of Clinical Sleep Medicine* **4**, 487–504 (2008).
- [64] D. Riemann, C. Baglioni, C. Bassetti, B. Bjorvatn, L. Dolenc Groselj, J. G. Ellis, C. A. Espie, D. Garcia-Borreguero, M. Gjerstad, M. Gonçalves, et al., "European guideline for the diagnosis and treatment of insomnia," *Journal of Sleep Research* **26**, 675–700 (2017).

- [65] F. S. Luyster, D. J. Buysse, and P. J. Strollo, "Comorbid insomnia and obstructive sleep apnea: challenges for clinical practice and research," *Journal of Clinical Sleep Medicine* **6**, 196–204 (2010).
- [66] R. R. Auger, H. J. Burgess, J. S. Emens, L. V. Deriy, S. M. Thomas, and K. M. Sharkey, "Clinical practice guideline for the treatment of intrinsic circadian rhythm sleep-wake disorders: advanced sleep-wake phase disorder (ASWPD), delayed sleep-wake phase disorder (DSWPD), non-24-hour sleep-wake rhythm disorder (N24SWD), and irregular sleep-wake rhythm disorder (ISWRD). An update for 2015," *Journal of Clinical Sleep Medicine* **11**, 1199–1236 (2015).
- [67] S. A. Rahman, L. Kayumov, E. A. Tchmoutina, and C. M. Shapiro, "Clinical efficacy of dim light melatonin onset testing in diagnosing delayed sleep phase syndrome," *Sleep Medicine* **10**, 549–555 (2009).
- [68] Z. Khan and L. M. Trotti, "Central disorders of hypersomnolence," *Chest* **148**, 262–273 (2015).
- [69] S. Overeem, E. Mignot, J. GertvanDijk, and G. J. Lammers, "Narcolepsy: clinical features, new pathophysiologic insights, and future perspectives," *Journal of Clinical Neurophysiology* **18**, 78–105 (2001).
- [70] P. Tinuper, F. Bisulli, and F. Provini, "The parasomnias: mechanisms and treatment," *Epilepsia* **53**, 12–19 (2012).
- [71] M. J. Howell, "Parasomnias: an updated review," *Neurotherapeutics* **9**, 753–775 (2012).
- [72] S. Abe, T. Yamaguchi, P. H. Rompre, P. De Grandmont, Y.-J. Chen, and G. J. Lavigne, "Tooth wear in young subjects: a discriminator between sleep bruxers and controls?," *International Journal of Prosthodontics* **22** (2009).
- [73] G. Fernandes, A. L. Franco, D. Aparecida de Godoi Gonçalves, J. Geraldo Speciali, M. E. Bigal, and C. M. Camparis, "Temporomandibular disorders, sleep bruxism, and primary headaches are mutually associated.," *Journal of Orofacial Pain* **27** (2013).
- [74] A. S. Walters, "Clinical identification of the simple sleep-related movement disorders," *Chest* **131**, 1260–1266 (2007).
- [75] G. Merlino and G. L. Gigli, "Sleep-related movement disorders," *Neurological Sciences* **33**, 491–513 (2012).
- [76] R. J. Broughton and J. M. Mullington, "Polysomnography: Principles and Applications in Sleep and Arousal Disorders," Chap 48 in *Electroencephalography : basic principles, clinical applications, and related fields*, Fifth edition. ed., E. Niedermeyer and F. L. d. Silva, eds. (Lippincott Williams & Wilkins, 2005).
- [77] E.-J. Speckmann and C. E. Elger, "Introduction to the Neurophysiological Basis of the EEG and DC Potentials," Chap 2 in *Electroencephalography : basic principles, clinical applications, and related fields*, Fifth edition. ed., E. Niedermeyer and F. L. d. Silva, eds. (Lippincott Williams & Wilkins, 2005).

- [78] E. O. Altenmüller, T. F. Münte, and C. Gerloff, "Neurocognitive Functions and the EEG," Chap 31 in *Electroencephalography : basic principles, clinical applications, and related fields*, Fifth edition. ed., E. Niedermeyer and F. L. d. Silva, eds. (Lippincott Williams & Wilkins, 2005).
- [79] H. H. Jasper, "The ten-twenty electrode system of the International Federation," *Electroencephalography and Clinical Neurophysiology* **10**, 370–375 (1958).
- [80] P. L. Nunez and R. Srinivasan, *Electric fields of the brain: the neurophysics of EEG* (Oxford University Press, USA, 2006).
- [81] A. Kamp, G. Pfurtscheller, G. Edlinger, and F. L. d. Silva, "Technological basis of EEG recording," Chap 6 in *Electroencephalography : basic principles, clinical applications, and related fields*, Fifth edition. ed., E. Niedermeyer and F. L. d. Silva, eds. (Lippincott Williams & Wilkins, 2005).
- [82] M. Weiergraeber, A. Papazoglou, K. Broich, and R. Mueller, "Sampling rate, signal bandwidth and related pitfalls in EEG analysis," *Journal of neuroscience methods* **268**, 53–55 (2016).
- [83] J. J. Halford, "Equipment and Instrumentation," Chap 2 in *Atlas of Artifacts in Clinical Neurophysiology*, W. O. Tatum IV, ed. (Springer Publishing Company, 2018).
- [84] A. Bulling, J. A. Ward, H. Gellersen, and G. Troster, "Eye movement analysis for activity recognition using electrooculography," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**, 741–753 (2010).
- [85] A. A. Alian and K. H. Shelley, "Photoplethysmography: Analysis of the pulse oximeter waveform," in *Monitoring Technologies in Acute Care Environments* (Springer, 2014), pp. 165–178.
- [86] P. D. Mannheimer, "The light–tissue interaction of pulse oximetry," *Anesthesia & Analgesia* **105**, S10–S17 (2007).
- [87] A. Delessert, F. Espa, A. Rossetti, G. Lavigne, M. Tafti, and R. Heinzer, "Pulse wave amplitude drops during sleep are reliable surrogate markers of changes in cortical activity," *Sleep* **33**, 1687–1692 (2010).
- [88] L. Grote, D. Zou, H. Kraiczi, and J. Hedner, "Finger plethysmography - A method for monitoring finger blood flow during sleep disordered breathing," *Respiratory Physiology and Neurobiology* **136**, 141–152 (2003).
- [89] M. Bruyneel, C. Sanida, G. Art, W. Libert, L. Cuvelier, M. Paesmans, R. Sergysels, and V. Ninane, "Sleep efficiency during sleep studies: Results of a prospective study comparing home-based and in-hospital polysomnography," *Journal of Sleep Research* **20**, 201–206 (2011).
- [90] R. Ferber, R. Millman, M. Coppola, J. Fleetham, C. Friederich Murray, C. Iber, W. V. McCall, G. Nino-Murcia, M. Pressman, M. Sanders, et al., "Portable recording in the assessment of obstructive sleep apnea," *Sleep* (1994).

- [91] J. Corral-Peñafiel, J.-L. Pepin, and F. Barbe, "Ambulatory monitoring in the diagnosis and management of obstructive sleep apnoea syndrome," *European Respiratory Review* **22**, 312–324 (2013).
- [92] P. Escourrou, L. Grote, T. Penzel, W. T. McNicholas, J. Verbraecken, R. Tkacova, R. L. Riha, J. Hedner, U. Anttalainen, F. Barbé, O. Basoglu, P. Bielicki, M. R. Bonsignore, C. Cerrato, R. Dumitrascu, C. Esquinas, D. Fairley, I. Fietze, A. Lopez, J. Kvamme, P. Levy, P. Sliwinski, M. Kumor, L. Lavie, P. Lavie, C. Lombardi, O. Marrone, J. F. Masa, J. M. Montserrat, G. Parati, J. Pépin, R. Plywaczewski, M. Pretl, S. Ryan, G. Roisman, T. Saaresranta, R. Schulz, P. Steiropoulos, S. Tasbakan, G. Varoneckas, H. Vrints, and M. Xanthoudaki, "The diagnostic method has a strong influence on classification of obstructive sleep apnea," *Journal of Sleep Research* **24**, 730–738 (2015).
- [93] L. Kalevo, T. Miettinen, A. Leino, S. Kainulainen, K. Myllymaa, J. Toyras, T. Leppanen, and S. Myllymaa, "Improved Sweat Artifact Tolerance of Screen-Printed EEG Electrodes by Material Selection-Comparison of Electrochemical Properties in Artificial Sweat," *IEEE Access* **7**, 133237–133247 (2019).
- [94] L. Kalevo, T. Miettinen, A. Leino, S. Kainulainen, H. Korkalainen, K. Myllymaa, J. Toyras, T. Leppanen, T. Laitinen, and S. Myllymaa, "Effect of Sweating on Electrode-Skin Contact Impedances and Artifacts in EEG Recordings With Various Screen-Printed Ag/AgCl Electrodes," *IEEE Access* **8**, 50934–50943 (2020).
- [95] T. Miettinen, K. Myllymaa, S. Westernen-Punnonen, J. Ahlberg, T. Hukkanen, J. Toyras, R. Lappalainen, E. Mervaala, K. Sipila, and S. Myllymaa, "Success Rate and Technical Quality of Home Polysomnography with Self-Applicable Electrode Set in Subjects with Possible Sleep Bruxism," *IEEE Journal of Biomedical and Health Informatics* **22**, 1124–1132 (2018).
- [96] D. J. Levendowski, L. Ferini-Strambi, C. Gamaldo, M. Cetel, R. Rosenberg, and P. R. Westbrook, "The accuracy, night-to-night variability, and stability of frontopolar sleep electroencephalography biomarkers," *Journal of Clinical Sleep Medicine* **13**, 791–803 (2017).
- [97] P. J. Arnal, V. Thorey, E. Debellemaniere, M. E. Ballard, A. Bou Hernandez, A. Guillot, H. Jourde, M. Harris, M. Guillard, P. Van Beers, M. Chennaoui, and F. Sauvet, "The Drem Headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging," *Sleep* (2020), zsa097.
- [98] A. Sadeh, "The role and validity of actigraphy in sleep medicine: An update," *Sleep Medicine Reviews* **15**, 259–267 (2011).
- [99] T. Morgenthaler, C. Alessi, L. Friedman, J. Owens, V. Kapur, B. Boehlecke, T. Brown, A. Chesson Jr, J. Coleman, T. Lee-Chiong, et al., "Practice parameters for the use of actigraphy in the assessment of sleep and sleep disorders: an update for 2007," *Sleep* **30**, 519–529 (2007).
- [100] J. Paquet, A. Kawinska, and J. Carrier, "Wake Detection Capacity of Actigraphy During Sleep," *Sleep* **30**, 1362–1369 (2007).

- [101] M. Marino, Y. Li, M. N. Rueschman, J. W. Winkelman, J. M. Ellenbogen, J. M. Solet, H. Dulin, L. F. Berkman, and O. M. Buxton, "Measuring Sleep: Accuracy, Sensitivity, and Specificity of Wrist Actigraphy Compared to Polysomnography," *Sleep* **36**, 1747–1755 (2013).
- [102] M. A. Nielsen, *Neural Networks and Deep Learning* (Determination Press, 2015).
- [103] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., "Human-level control through deep reinforcement learning," *Nature* **518**, 529–533 (2015).
- [104] F. Rosenblatt, *Principles of neurodynamics. perceptrons and the theory of brain mechanisms* (Cornell Aeronautical Lab Inc Buffalo NY, 1961).
- [105] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400* (2013).
- [106] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation* **9**, 1735–1780 (1997).
- [107] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259* (2014).
- [108] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis* **42**, 60–88 (2017).
- [109] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine* **25**, 65 (2019).
- [110] Z. I. Attia, S. Kapa, F. Lopez-Jimenez, P. M. McKie, D. J. Ladewig, G. Satam, P. A. Pellikka, M. Enriquez-Sarano, P. A. Noseworthy, T. M. Munger, et al., "Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram," *Nature Medicine* **25**, 70–74 (2019).
- [111] C. A. Goldstein, R. B. Berry, D. T. Kent, D. A. Kristo, A. A. Seixas, S. Redline, M. B. Westover, F. Abbasi-Feinberg, R. N. Aurora, K. A. Carden, et al., "Artificial intelligence in sleep medicine: an American Academy of Sleep Medicine position statement," *Journal of Clinical Sleep Medicine* **16**, 605–607 (2020).
- [112] S. Nikkonen, I. O. Afara, T. Leppänen, and J. Töyräs, "Artificial neural network analysis of the oxygen saturation signal enables accurate diagnostics of sleep apnea," *Scientific Reports* **1-9** (2019).
- [113] G. C. Gutiérrez-Tobal, D. Álvarez, A. Crespo, F. Del Campo, and R. Hornero, "Evaluation of machine-learning approaches to estimate sleep apnea severity from at-home oximetry recordings," *IEEE Journal of Biomedical and Health Informatics* **23**, 882–892 (2018).

- [114] F. Vaquerizo-Villar, D. Álvarez, L. Kheirandish-Gozal, G. C. Gutiérrez-Tobal, V. Barroso-García, F. del Campo, D. Gozal, and R. Hornero, "Convolutional Neural Networks to Detect Pediatric Apnea-Hypopnea Events from Oximetry," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE, 2019), pp. 3555–3558.
- [115] J. B. Stephansen, A. N. Olesen, M. Olsen, A. Ambati, E. B. Leary, H. E. Moore, O. Carrillo, L. Lin, F. Han, H. Yan, et al., "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy," *Nature Communications* **9**, 1–15 (2018).
- [116] M. Shahin, B. Ahmed, S. T.-B. Hamida, F. L. Mulaffer, M. Glos, and T. Penzel, "Deep learning and insomnia: Assisting clinicians with their diagnosis," *IEEE Journal of Biomedical and Health Informatics* **21**, 1546–1553 (2017).
- [117] A. V. Zinchuk, S. Jeon, B. B. Koo, X. Yan, D. M. Bravata, L. Qin, B. J. Selim, K. P. Strohl, N. S. Redeker, J. Concato, and H. K. Yaggi, "Polysomnographic phenotypes and their cardiovascular implications in obstructive sleep apnoea," *Thorax* **73**, 472 (2018).
- [118] P. Anderer, G. Gruber, S. Parapatics, M. Woertz, T. Miazhyńskaia, G. Klösch, B. Saletu, J. Zeitlhofer, M. J. Barbanoj, H. Danker-Hopfe, S. L. Himanen, B. Kemp, T. Penzel, M. Grözinger, D. Kunz, P. Rappelsberger, A. Schlögl, and G. Dorffner, "An E-Health solution for automatic sleep classification according to Rechtschaffen and Kales: Validation study of the somnolyzer 24 x 7 utilizing the siesta database," *Neuropsychobiology* **51**, 115–133 (2005).
- [119] B. Koley and D. Dey, "An ensemble system for automatic sleep stage classification using single channel EEG signal," *Computers in Biology and Medicine* **42**, 1186–1195 (2012).
- [120] C. Stepnowsky, D. Levendowski, D. Popovic, I. Ayappa, and D. M. Rapoport, "Scoring accuracy of automated sleep staging from a bipolar electroocular recording compared to manual scoring by multiple raters," *Sleep Medicine* **14**, 1199–1207 (2013).
- [121] S. Mousavi, F. Afghah, and U. R. Acharya, "SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PloS one* **14**, e0216456 (2019).
- [122] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **25**, 1998–2008 (2017).
- [123] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. D. Vos, "Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification," *IEEE Transactions on Biomedical Engineering* **66**, 1285–1296 (2019).
- [124] O. Tsinalis, P. M. Matthews, and Y. Guo, "Automatic Sleep Stage Scoring Using Time-Frequency Analysis and Stacked Sparse Autoencoders," *Annals of Biomedical Engineering* **44**, 1587–1597 (2016).

- [125] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, "Automatic Sleep Stage Scoring with Single-Channel EEG Using Convolutional Neural Networks," *arXiv preprint arXiv:1610.01683* (2016).
- [126] H. Phan, F. Andreotti, N. Cooray, O. Chén, and M. de Vos, "SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging," *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **27**, 400–410 (2019).
- [127] F. Andreotti, H. Phan, N. Cooray, C. Lo, M. T. Hu, and M. D. Vos, "Multichannel Sleep Stage Classification and Transfer Learning using Convolutional Neural Networks," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2018-July*, 171–174 (2018).
- [128] A. Patanaik, J. L. Ong, J. J. Gooley, S. Ancoli-Israel, and M. W. Chee, "An end-to-end framework for real-time automatic sleep stage classification," *Sleep* **41**, 1–11 (2018).
- [129] H. Sun, J. Jia, B. Goparaju, G.-B. Huang, O. Sourina, M. T. Bianchi, and M. B. Westover, "Large-Scale Automated Sleep Staging," *Sleep* **40** (2017).
- [130] S. Biswal, H. Sun, B. Goparaju, M. B. Westover, J. Sun, and M. T. Bianchi, "Expert-level sleep scoring with deep neural networks," *Journal of the American Medical Informatics Association* **25**, 1643–1650 (2018).
- [131] A. Malafeev, D. Laptev, S. Bauer, X. Omlin, A. Wierzbicka, A. Wichniak, W. Jernajczyk, R. Riener, J. Buhmann, and P. Achermann, "Automatic Human Sleep Stage Scoring Using Deep Neural Networks.," *Frontiers in Neuroscience* **12**, 781 (2018).
- [132] H. Phan, K. Mikkelsen, O. Y. Chén, P. Koch, A. Mertins, P. Kidmose, and M. De Vos, "Personalized automatic sleep staging with single-night data: a pilot study with KL-divergence regularization," *Physiological Measurement* (2020).
- [133] G. Zhu, Y. Li, and P. P. Wen, "Analysis and classification of sleep stages based on difference visibility graphs from a single-channel EEG signal," *IEEE Journal of Biomedical and Health Informatics* **18**, 1813–1821 (2014).
- [134] Q. Li, Q. Li, C. Liu, S. P. Shashikumar, S. Nemati, and G. D. Clifford, "Deep learning in the cross-time frequency domain for sleep staging from a single-lead electrocardiogram," *Physiological Measurement* **39**, 124005 (2018).
- [135] P. Fonseca, X. Long, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink, "Sleep stage classification with ECG and respiratory effort," *Physiological Measurement* **36**, 2027–2040 (2015).
- [136] T. Willemen, D. V. Deun, V. Verhaert, M. Vandekerckhove, V. Exadaktylos, J. Verbraecken, S. V. Huffel, B. Haex, and J. V. Sloten, "An evaluation of cardiorespiratory and movement features with respect to sleep-stage classification," *IEEE Journal of Biomedical and Health Informatics* **18**, 661–669 (2014).

- [137] P. Fonseca, M. M. van Gilst, M. Radha, M. Ross, A. Moreau, A. Cerny, P. Anderer, X. Long, J. P. van Dijk, and S. Overeem, "Automatic sleep staging using heart rate variability, body movements, and recurrent neural networks in a sleep disordered population," *Sleep* (2020), zsa048.
- [138] P. Fonseca, T. Weysen, M. S. Goelema, E. I. Møst, M. Radha, C. L. Scheurleer, L. V. D. Heuvel, and R. M. Aarts, "Validation of photoplethysmography-based sleep staging compared with polysomnography in healthy middle-aged adults," *Sleep* **40** (2017).
- [139] Z. Beattie, Y. Oyang, A. Statan, A. Ghoreyshi, A. Pantelopoulos, A. Russell, and C. Heneghan, "Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals," *Physiological Measurement* **38**, 1968–1979 (2017).
- [140] M. K. Uçar, M. R. Bozkurt, C. Bilgin, and K. Polat, "Automatic sleep staging in obstructive sleep apnea patients using photoplethysmography, heart rate variability signal and machine learning techniques," *Neural Computing and Applications* **29**, 1–16 (2018).
- [141] P. Dehkordi, A. Garde, G. A. Dumont, and J. M. Ansermino, "Sleep/wake classification using cardiorespiratory features extracted from photoplethysmogram," *Computing in Cardiology* **43**, 1021–1024 (2016).
- [142] M. A. Motin, C. K. Karmakar, T. Penzel, and M. Palaniswami, "Sleep-Wake Classification using Statistical Features Extracted from Photoplethysmographic Signals," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* **c**, 5564–5567 (2019).
- [143] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. Kamphuisen, and J. J. Oberyé, "Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG," *IEEE Transactions on Biomedical Engineering* **47**, 1185–1194 (2000).
- [144] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation* **101**, e220 (2000).
- [145] C. O'reilly, N. Gosselin, J. Carrier, and T. Nielsen, "Montreal Archive of Sleep Studies: an open-access resource for instrument benchmarking and exploratory research," *Journal of sleep research* **23**, 628–635 (2014).
- [146] M. Younes, M. Ostrowski, M. Soiferman, H. Younes, M. Younes, J. Raneri, and P. Hanly, "Odds Ratio Product of Sleep EEG as a Continuous Measure of Sleep State," *Sleep* **38**, 641–654 (2015).
- [147] P. Tiisonen, *Novel portable devices for recording sleep apnea and evaluation altered consciousness*, (vältöskirja, Kuopion yliopisto, 2009).
- [148] T. Leppänen, J. Töyräs, E. Mervaala, T. Penzel, and A. Kulkas, "Severity of individual obstruction events increases with age in patients with obstructive sleep apnea," *Sleep Medicine* **37**, 32–37 (2017).

- [149] D. R. Cox, "Regression Models and Life-Tables," *Journal of the Royal Statistical Society. Series B (Methodological)* **34**, 187–220 (1972).
- [150] D. R. Cox, "Partial Likelihood," *Biometrika* **62** (1975).
- [151] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *arXiv preprint arXiv:1409.0473* (2014).
- [152] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," *arXiv preprint arXiv:1412.3555* (2014).
- [153] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," *arXiv preprint arXiv:1608.03983* 1-16 (2016).
- [154] L. N. Smith, "Cyclical learning rates for training neural networks," *Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017* 464-472 (2017).
- [155] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement* **20**, 37–46 (1960).
- [156] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
- [157] A. Kulkas, P. Tiihonen, P. Julkunen, E. Mervaala, and J. Töyräs, "Novel parameters indicate significant differences in severity of obstructive sleep apnea with patients having similar apnea-hypopnea index," *Medical & Biological Engineering & Computing* **51**, 697–708 (2013).
- [158] D. W. Hudgel, "Sleep Apnea Severity Classification - Revisited," *Sleep* **39**, 1165–1166 (2016).
- [159] H. Danker-Hopfe, P. Anderer, J. Zeitlhofer, M. Boeck, H. Dorn, G. Gruber, E. Heller, E. Loretz, D. Moser, S. Parapatics, B. Saletu, A. Schmidt, and G. Dorffner, "Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard," *Journal of Sleep Research* **18**, 74–84 (2009).
- [160] U. J. Magalang, N.-H. Chen, P. A. Cistulli, A. C. Fedson, T. Gíslason, D. Hillman, T. Penzel, R. Tamisier, S. Tufik, G. Phillips, and A. I. Pack, "Agreement in the Scoring of Respiratory Events and Sleep Among International Sleep Centers," *Sleep* **36**, 591–596 (2013).
- [161] X. Zhang, X. Dong, J. W. Kantelhardt, J. Li, L. Zhao, C. Garcia, M. Glos, T. Penzel, and F. Han, "Process and outcome for international reliability in sleep scoring," *Sleep and Breathing* **19**, 191–195 (2015).
- [162] B. Duce, C. Rego, J. Milosavljevic, and C. Hukins, "The AASM recommended and acceptable EEG montages are comparable for the staging of sleep and scoring of EEG arousals," *Journal of Clinical Sleep Medicine* **10**, 803–809 (2014).

- [163] T. Penzel, X. Zhang, and I. Fietze, "Inter-scoring Reliability between Sleep Centers Can Teach Us What to Improve in the Scoring Rules," *Journal of Clinical Sleep Medicine* **9**, 89 (2013).
- [164] 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* **23**, 901–908 (2000).
- [165] R. S. Rosenberg and S. V. Hout, "The American Academy of Sleep Medicine inter-scoring reliability program: Respiratory events," *Journal of Clinical Sleep Medicine* **10**, 447–454 (2014).
- [166] S. Myllymaa, A. Muraja-Murro, S. Westeren-Punnonen, T. Hukkanen, R. Lappalainen, E. Mervaala, J. Töyräs, K. Sipilä, and K. Myllymaa, "Assessment of the suitability of using a forehead EEG electrode set and chin EMG electrodes for sleep staging in polysomnography.," *Journal of Sleep Research* **25**, 636–645 (2016).
- [167] M. de Zambotti, L. Rosas, I. M. Colrain, and F. C. Baker, "The Sleep of the Ring: Comparison of the ÖURA Sleep Tracker Against Polysomnography," *Behavioral Sleep Medicine* **17**, 124–136 (2019).
- [168] M. D. Zambotti, N. Cellini, A. Goldstone, I. M. Colrain, and F. C. Baker, "Wearable Sleep Technology in Clinical and Research Settings," *Medicine and Science in Sports and Exercise* **51**, 1538–1557 (2019).
- [169] K. R. Evenson, M. M. Goto, and R. D. Furberg, "Systematic review of the validity and reliability of consumer-wearable activity trackers," *International Journal of Behavioral Nutrition and Physical Activity* **12** (2015).
- [170] A. Gruwez, A. V. Bruyneel, and M. Bruyneel, "The validity of two commercially-available sleep trackers and actigraphy for assessment of sleep parameters in obstructive sleep apnea patients," *PLoS ONE* **14**, 1–11 (2019).
- [171] M. H. Asyali, R. B. Berry, M. C. Khoo, and A. Altinok, "Determining a continuous marker for sleep depth," *Computers in Biology and Medicine* **37**, 1600–1609 (2007).
- [172] M. J. Drinnan, A. Murray, C. J. Griffiths, and G. J. Gibson, "Interobserver Variability in Recognizing Arousal in Respiratory Sleep Disorders," *American Journal of Respiratory and Critical Care Medicine* **158**, 358–362 (1998).

ORIGINAL PUBLICATIONS (I-IV)

Paper II

Korkalainen H, Aakko J, Nikkonen S, Kainulainen S, Leino A, Duce B, Afara IO, Myllymaa S, Töyräs J, and Leppänen T.

Accurate Deep Learning-Based Sleep Staging in a Clinical Population with Suspected Obstructive Sleep Apnea











IEEE Journal of Biomedical and Health Informatics,

24(7): 2073-2081, 2019.

DOI: 10.1109/JBHI.2019.2951346

Reprinted under the the terms of Creative Commons Attribution License
(CC BY 4.0)

Accurate Deep Learning-Based Sleep Staging in a Clinical Population With Suspected Obstructive Sleep Apnea

Henri Korkalainen , Juhani Aakko , Sami Nikkonen , Samu Kainulainen , Akseli Leino , Brett Duce , Isaac O. Afara , Sami Myllymaa , Juha Töyräs , and Timo Leppänen 

Abstract—The identification of sleep stages is essential in the diagnostics of sleep disorders, among which obstructive sleep apnea (OSA) is one of the most prevalent. However, manual scoring of sleep stages is time-consuming, subjective, and costly. To overcome this shortcoming, we aimed to develop an accurate deep learning approach for automatic classification of sleep stages and to study the effect of OSA severity on the classification accuracy. Overnight polysomnographic recordings from a public dataset of healthy individuals (Sleep-EDF, $n = 153$) and from a clinical dataset ($n = 891$) of patients with

suspected OSA were used to develop a combined convolutional and long short-term memory neural network. On the public dataset, the model achieved sleep staging accuracy of 83.7% ($\kappa = 0.77$) with a single frontal EEG channel and 83.9% ($\kappa = 0.78$) when supplemented with EOG. For the clinical dataset, the model achieved accuracies of 82.9% ($\kappa = 0.77$) and 83.8% ($\kappa = 0.78$) with a single EEG channel and two channels (EEG + EOG), respectively. The sleep staging accuracy decreased with increasing OSA severity. The single-channel accuracy ranged from 84.5% ($\kappa = 0.79$) for individuals without OSA diagnosis to 76.5% ($\kappa = 0.68$) for patients with severe OSA. In conclusion, deep learning enables automatic sleep staging for suspected OSA patients with high accuracy and expectedly, the accuracy decreased with increasing OSA severity. Furthermore, the accuracies achieved in the public dataset were superior to previously published state-of-the-art methods. Adding an EOG channel did not significantly increase the accuracy. The automatic, single-channel-based sleep staging could enable easy, accurate, and cost-efficient integration of EEG recording into diagnostic ambulatory recordings.

Manuscript received May 27, 2019; revised September 17, 2019 and October 30, 2019; accepted October 31, 2019. This work was supported in part by the Research Committee of the Kuopio University Hospital Catchment Area for the State Research Funding (Projects 5041767, 5041768, 5041770, 5041776, 5041779, 5041780, 5041781, and 5041783), in part by the Academy of Finland (decisions 313697 and 323536), in part by the Respiratory Foundation of Kuopio Region, in part by the Research Foundation of the Pulmonary Diseases, in part by Foundation of the Finnish Anti-Tuberculosis Association, in part by the Päivikki and Sakari Sohlberg Foundation, in part by Orion Research Foundation, in part by Instrumentarium Science Foundation, in part by the Finnish Cultural Foundation via the Post Docs in Companies program and via the Central Fund, in part by the Paulo Foundation, in part by the Tampere Tuberculosis Foundation, and in part by Business Finland (decision 5133/31/2018). (Corresponding author: Henri Korkalainen.)

H. Korkalainen, S. Nikkonen, S. Kainulainen, A. Leino, S. Myllymaa, and T. Leppänen are with the Department of Applied Physics, University of Eastern Finland, Kuopio 70210, Finland, and also with the Diagnostic Imaging Center, Kuopio University Hospital, Kuopio 70210, Finland (e-mail: henri.korkalainen@uef.fi; sami.nikkonen@uef.fi; samu.kainulainen@uef.fi; akseli.leino@uef.fi; sami.myllymaa@uef.fi; timo.leppanen@uef.fi).

J. Aakko is with the CGI Suomi Oy, Helsinki 00380, Finland (e-mail: juhani.aakko@cgi.com).

B. Duce is with the Sleep Disorders Centre, Department of Respiratory & Sleep Medicine, Princess Alexandra Hospital, Woolloongabba, QLD 4102, Australia, and also with the Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane City, QLD 4000, Australia (e-mail: brett.duce@health.qld.gov.au).

I. O. Afara is with the Department of Applied Physics, University of Eastern Finland, Kuopio 70210, Finland, and also with the School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, QLD 4072, Australia (e-mail: isaac.afara@uef.fi).

J. Töyräs is with the Department of Applied Physics, University of Eastern Finland, Kuopio 70210, Finland, with the Diagnostic Imaging Center, Kuopio University Hospital, Kuopio 70210, Finland, and also with the School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, QLD 4072, Australia (e-mail: juha.toyras@uef.fi).

Digital Object Identifier 10.1109/JBHI.2019.2951346

Index Terms—Deep learning, Electroencephalography, Obstructive sleep apnea, Recurrent neural network, Sleep staging.

1. INTRODUCTION

IDENTIFICATION of sleep stages is crucial in diagnostics of various sleep disorders. One of the most common sleep disorders is obstructive sleep apnea (OSA) which has been estimated to affect up to 38% of the general population [1]. In the diagnosis of OSA, sleep staging is conducted to assess the sleep characteristics and to accurately determine the total sleep time [2]. Accurate determination of total sleep time is of paramount importance as it significantly affects the parameters used to assess the severity of OSA.

According to the current sleep staging criteria [2], sleep is classified into five different stages: wake, rapid eye movement (REM) sleep and three stages of non-REM sleep (N1–N3). Classification into these stages is performed manually for 30-second epochs of sleep using electroencephalography (EEG), electrooculogram (EOG), and submental electromyogram (EMG) signals measured during polysomnography (PSG). Currently, at least 13 electrodes, with the positions determined by the International 10-20 System, are required for the measurement protocol [2]. Thus, the overall measurement protocol and the sleep staging

process is time-consuming, laborious, and requires experienced professionals [3].

Despite the major effort and expenses that go into manual sleep staging, there are still shortcomings. Mainly, the agreement of two different scorers is generally unsatisfactory [4]–[9]. The inter-rater reliability (IRR), measured with Cohen’s kappa, between two scorers using the current sleep scoring criteria is commonly around 0.78 [4]. However, between international sleep centers, the reliability can be as low as 0.58 to 0.63 [5], [6], particularly due to poor scoring of N1 sleep [7], [8]. It has been shown that the agreement of N1 is approximately only 0.46 between sleep laboratories within Europe [4] and as low as 0.19 to 0.31 between international centers [5], [6]. Furthermore, the overall reliability of manual sleep staging may further decrease if an individual is experiencing medical conditions, for example, with OSA patients the reliability is worse than that of healthy individuals [8], [9]. Automatic scoring methods could potentially improve the consistency of sleep staging between different hospitals and healthcare systems. Furthermore, automatic methods capable of accurate sleep staging with a minimal number of measured signals could simplify the measurement protocol and reduce the related costs.

A number of automatic sleep staging methods have been previously published [10]–[31]. Traditionally, automated methods have relied on pre-defined rules, carefully selected features extracted from the signals, and classification algorithms [22]–[26]. Recently, a few machine-learning-based solutions utilizing deep learning and artificial neural networks have been presented [10]–[12], [14], [16]–[21], [27]–[31]. For these solutions, the classification rules or features of each sleep stage were not explicitly defined. However, previous studies have generally relied on heavy preprocessing by either transforming the signals into 2D images representing the spectral information [19], [27]–[30] or by reducing the signals into a limited number of predefined features [10], [30], [31]. Furthermore, deep learning models developed on research datasets of healthy individuals have generally suffered from a loss of accuracy when generalizing into populations with sleep disorders such as OSA [28]. In addition, a few machine learning-based automation attempts have demonstrated promising outcomes on sleep staging with a single EEG channel [10], [11], [13]–[16], [18]–[21]. While some of these have utilized deep learning [10], [11], [14], [16], [18], [19], [21], they have mostly relied on publicly available research datasets with a limited number of healthy individuals. Large clinical and well-balanced datasets have rarely been used, and the effect of sleep disorders on automatic sleep staging has not been thoroughly investigated.

We aimed to develop an accurate deep learning-based automatic method for the classification of sleep stages in patients with suspected OSA. We further aimed to achieve this by utilizing the raw signals without conducting heavy preprocessing. Furthermore, we aimed to study the effect of OSA severity on the performance of automatic sleep staging. We hypothesize that deep learning methods enable accurate sleep staging based on a single EEG channel for patients with suspected OSA and that the sleep staging accuracy decreases with increasing OSA severity.

II. METHODS

A. Datasets

1) *Sleep-EDF*: We first utilized a public dataset, Physionet Sleep-EDF [32], [33], to allow comparison of the proposed deep learning-based approach with previous state-of-the-art methods. We utilized the version 2 of the expanded Sleep-EDF dataset released in March 2018. The dataset comprises 153 PSGs of 37 males and 41 females from a study investigating the effects of age on sleep in a healthy population (Sleep Cassette). We utilized the Fpz-Cz EEG signal for a single-channel input and combined it with a single horizontal EOG signal for two-channel input. Both signals were sampled with a 100 Hz frequency. No preprocessing was implemented on the signals. EMG recording was left out of this study due to its lower sampling frequency.

The sleep stages were originally scored according to the Rechtschaffen and Kales manual [34] into following stages: wake, N1, N2, N3, N4, REM, M (movement), and ‘?’ (not scored). We combined the stages N3 and N4 into a single sleep stage to comply with the AASM guidelines [2]. Furthermore, the stages M and ‘?’ were excluded from the study. The PSG recordings included long periods of wake in the beginning and end of the recording. Similarly to previous studies [11], [18], we only included 30 minutes of the wake before and after the sleep to obtain more realistic results and to enable comparison.

With the Sleep-EDF dataset, we conducted 10-fold cross-validation to assess the performance of the network, meaning that with each fold, 90% of the population was used for training and 10% as an independent test set. Furthermore, 10% of the training set was further used as the validation set during each fold. This was done to avoid overfitting during training, to choose an optimal model, and to keep the test set separate during each fold. 10-fold cross-validation was chosen over a single split to training, validation, and test set due to relatively small dataset and to enable comparison with the previous studies [11], [17]–[21].

2) *Clinical Dataset*: The clinical dataset utilized in this study consists of 933 consecutive diagnostic overnight polysomnographies (PSG) of patients with clinical suspicion of OSA. Out of these, 891 individuals had successful recordings of all the required signals together with complete sleep stage scorings and were thus included in this study. The PSGs were conducted at the Princess Alexandra Hospital, Brisbane, Australia during 2015–2017 and recorded with the Compumedics Graef acquisition system (Compumedics, Abbotsford, Australia). The sleep stages were initially scored manually by multiple experienced scorers who participate regularly in intra- and inter-laboratory scoring concordance activities. Scoring was conducted based on the AASM rules [2] and the prevailing clinical practice of the Princess Alexandra Hospital. Ethical permissions for the data collection and processing were obtained from The Institutional Human Research Ethics Committee of the Princess Alexandra Hospital (HREC/16/QPAH/021 and LNR/2019/QMS/54313).

From the recorded PSGs, EEG (derivation F4-M1) was used for single-channel input and it was complemented with EOG (derivation E1-M2) for two-channel input. EMG was not included to enable comparison with the public dataset. The signals

TABLE I
DEMOGRAPHIC INFORMATION OF THE CLINICAL DATASET ($n = 891$)

	Median	Lower and upper quartiles
Apnea-hypopnea index (events/hour)	15.8	7.0–32.8
Age (years)	55.8	44.7–65.8
Body mass index (kg/m ²)	34.5	29.4–40.4
Arousal index (arousals/hour)	20.8	14.0–31.4
Total recording time (min)	442.5	409.5–474.5
Total sleep time (min)	308.8	253.8–359.8
Wake after sleep onset (min)	102.8	61.3–150
Sleep latency (min)	17.5	9.0–34.5
N1 (%)	11.0	6.8–18.9
N2 (%)	48.3	41.3–56.2
N3 (%)	18.3	9.6–27.1
REM (%)	17.1	11.8–22.1
NREM (%)	82.9	77.8–88.1
Sleep efficiency (%)	70.7	57.9–82.0

N1, N2, N3, and REM mean the percentage of the sleep stage and NREM the percentage of non-REM sleep from total sleep time. Sleep efficiency means the percentage of sleep from total recording time.

were recorded with 1024 Hz sampling frequency and were downsampled to 64 Hz to reduce the computational load. No additional preprocessing was applied. The frontal EEG channel was selected due to its simple measurement setup. The dataset was split into three individual sets: a training set, a validation set, and a test set. The training set comprised 717 whole night recordings (80%), and the validation and test sets comprised 87 recordings (10%) each.

Out of the 891 studied individuals, 493 were males and 398 females. The patients were mostly middle-aged and obese. According to the current severity classification of OSA, based on apnea-hypopnea index (AHI) [35], 152 individuals had no OSA ($5 < \text{AHI}$), 278 suffered from mild OSA ($5 \leq \text{AHI} < 15$), 208 from moderate OSA ($15 \leq \text{AHI} < 30$), and 254 had severe OSA ($\text{AHI} \geq 30$). Furthermore, 142 individuals were smokers, 197 suffered from diabetes, 368 had hypertension, 96 had cardiac arrhythmia, 22 had cardiac failure, and 41 had suffered a stroke. **Table I** shows the medians and interquartile ranges for sleep parameters and demographic information.

3) OSA Severity: The effect of OSA severity on the performance of the automatic sleep staging model was assessed by training and evaluating the model separately on each OSA severity group (no OSA, mild, moderate, and severe OSA) of the clinical dataset described above. In this phase, only a single frontal EEG channel (F4-M1) was used, and as with the Sleep-EDF dataset, the performance was evaluated using 10-fold cross-validation. The 10-fold cross-validation was chosen due to reduced size of the dataset compared to the complete clinical dataset, and to get more comprehensive and comparable results over all the severity groups. **Table II** presents the number of 30-second epochs of each sleep stage in all the utilized datasets.

B. Neural Network Architecture

The estimation of the sleep stages (wake, N1, N2, N3, and REM) was conducted with a combined convolutional network

(CNN) and recurrent neural network (RNN) trained in an end-to-end manner. The CNN aspect of the network was used to learn the characteristic features typical of each sleep stage, while the RNN considered the temporal distribution of the sleep stages overnight. The combined CNN and RNN structure was in essence similar to the architecture presented earlier by Supratak *et al.* [11]. However, sleep staging was conducted as a sequence-to-sequence classification problem, previously proposed by Phan *et al.* [29]. The network architecture was identical for the two-channel input and the single-channel input; the only difference was in the input dimension. The network was implemented in Python 3.6 using Keras API 2.2.4 with TensorFlow (version 1.13) backend. The training was conducted on a server with 32-core AMD Ryzen Threadripper 2990WX, 128 GB RAM and NVIDIA GeForce RTX 2080.

The CNN comprised six 1D convolutions each followed by batch normalization and a rectified linear unit (ReLU) activation, two max-pooling layers, and a global average pooling layer (**Fig. 1**). The max-pooling layers were situated after the first two 1D convolutions and after the two following 1D convolutions. The global average pooling layer followed the last two 1D convolutions. The kernel size of the first 1D convolution was 21 and the stride size was 5. The second 1D convolution had a kernel size of 21 and stride size of 1. The number of convolutional filters equaled the sampling frequency (64 Hz for the clinical dataset, 100 Hz for Sleep-EDF) of the used dataset in the first two 1D convolutions. The remaining four 1D convolutions had a kernel size of 5 with a stride size of 1. The number of convolutional filters was two times the sampling frequency for the third and fourth 1D convolution and four times the frequency for the fifth and sixth 1D convolution.

The complete network comprised a time distributed layer of the complete CNN structure, a gaussian dropout layer and a bidirectional long short-term memory (LSTM) layer followed by time distributed dense layer with softmax activation (**Fig. 1**). The number of units in the bidirectional LSTM was 4 times the sampling frequency. The LSTM utilized a tanh activation function and a dropout rate of 0.3. In the recurrent step, a hard sigmoid activation and a dropout rate of 0.5 were used. The last layer of the network comprised a dense layer with softmax activation producing the output sequence of sleep stage probabilities.

The model was trained with sequences of hundred 30-second epochs. An overlap of 75% was used when forming the sequences in the training set to increase its size fourfold. No overlap was used in the validation set or the test set. The model was trained with categorical cross-entropy as the loss function and an Adam optimizer with warm restarts [36] using a learning rate range of 0.001 to 0.00001. This learning rate range was optimized with a learning rate finder [37]. The model was validated with the validation set after each training cycle *i.e.*, after the entire training set was passed through the network. The model was trained for a maximum of 200 training cycles or until the value of the loss function in the validation set no longer decreased during 20 consecutive training cycles. The performance of the model was then assessed using in an independent test set.

TABLE II
THE NUMBER OF 30-SECOND EPOCHS OF EACH SLEEP STAGE IN THE SLEEP-EDF ATASET, CLINICAL DATASET, AND AMONG THE GROUPS WITH DIFFERENT OSAS SEVERITY

	Wake	N1	N2	N3	REM	Total
Sleep-EDF	65655 (34%)	21522 (11%)	69132 (35%)	13039 (7%)	25835 (13%)	195183
Clinical dataset	254278 (32%)	74102 (9%)	261317 (33%)	105298 (13%)	95800 (12%)	790795
No OSA	37303 (28%)	7501 (6%)	47782 (35%)	23076 (17%)	19262 (14%)	134924
Mild OSA	70532 (29%)	17947 (7%)	88412 (36%)	37554 (15%)	32485 (13%)	246930
Moderate OSA	59653 (32%)	15938 (9%)	61534 (33%)	25340 (14%)	22820 (12%)	185285
Severe OSA	86790 (39%)	32716 (15%)	63589 (28%)	19328 (9%)	212339 (9%)	223656

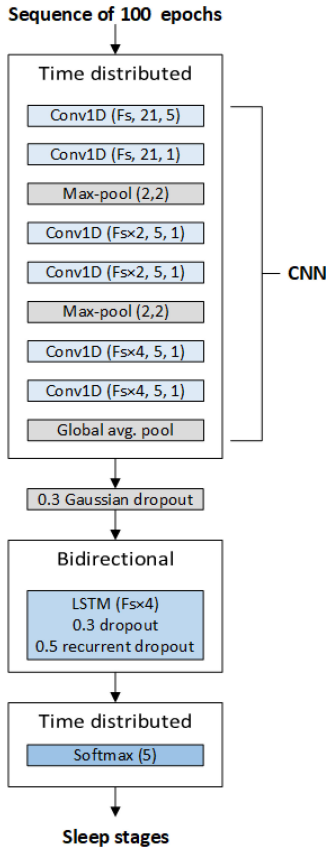


Fig. 1. The architecture of the combined convolutional neural network (CNN) and recurrent neural network (RNN). The parameters of the 1D convolutions (Conv1D) are given as (number of filters, kernel size, stride size) and as (pool size, stride size) for the max-pooling. F_s is the sampling frequency. For the long short-term memory (LSTM) and dense layer (Softmax) the number of units is given. The rate is given for the dropouts. The dropout layers were only activated during training. Sequences of hundred 30-second epochs of the utilized signals were used as an input, and the model produced a sequence of softmax values representing the probabilities of each sleep stage for each epoch.

C. Interpretation of the Results

The accuracies were calculated in an epoch-by-epoch manner. Moreover, the inter-rater agreement between manual and automatic sleep staging was evaluated using Cohen's kappa coefficient (κ) [38] and the sensitivity and specificity of differentiating sleep from wake sleep were calculated.

III. RESULTS

A. Sleep-EDF

During the 10-fold cross-validation, the model achieved 89.8% training accuracy, 83.0% validation accuracy, and 83.9% testing accuracy with the two-channel input comprising single EEG and EOG channels. These accuracies corresponded to kappa values of 0.86, 0.77, and 0.78 in the training, validation, and test sets, respectively. Based on the guidelines by Landis and Koch [39], the kappa values indicate almost perfect agreement between manual and automatic sleep staging in the training set, and substantial agreement in the validation and test sets. In the test set, sleep was identified with 96.2% sensitivity and 93.7% specificity. For the individual sleep stages, the accuracy was 93.7% for wake, 87.3% for N2, 78.0% for N3, and 85.4% for REM in the test sets, Fig. 2A. The lowest concordance was seen with N1 (45.1%).

With the single EEG channel, the obtained accuracies were 89.2%, 82.8%, and 83.7% in training, validation, and test sets, respectively. These correspond to kappa values of 0.85, 0.77, 0.77, respectively, indicating almost perfect or substantial agreement. In the test set, sleep was identified with 96.0% sensitivity and 93.4% specificity. Wake was identified with 93.4%, N1 with 43.4%, N2 with 87.3%, N3 with 78.7%, and REM with 85.4% accuracy (Fig. 2B). The obtained accuracies and kappa values with single and two-channel input, alongside previous state-of-the-art results, are presented in Table III.

B. Clinical Dataset

In the clinical dataset with the F4-M1 EEG and E1-M2 EOG channels, the model achieved 85.5% training accuracy and 83.8% validation accuracy. In the independent test set, the accuracy was 83.8%. These accuracies corresponded to Cohen's kappa values of 0.80, 0.78, and 0.78, respectively, indicating substantial agreement. Furthermore, the sensitivity of identifying sleep was 95.9% with 89.4% specificity in the test set. For individual sleep stages, the accuracy was 89.4% for wake, 87.2% for N2, 79.8% for N3 and 91.4% for REM in the test set (Fig. 3A). The lowest concordance between manual and automatic sleep staging was obtained in N1 with an accuracy of 46.9%.

With the single frontal EEG channel, the accuracies were 86.3%, 83.4%, and 82.9% in the training, validation and test

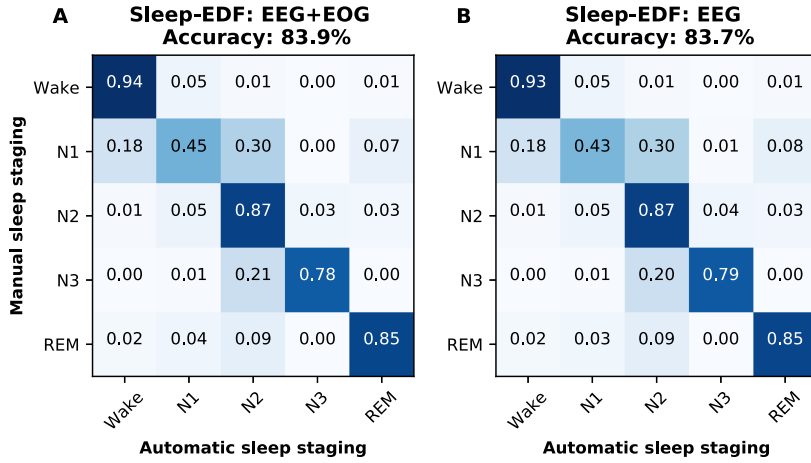


Fig. 2. Normalized confusion matrices of the classification accuracies from Sleep-EDF with (A) two-channel input (Fpz-Cz EEG and horizontal EOG) and (B) single EEG channel (Fpz-Cz) input.

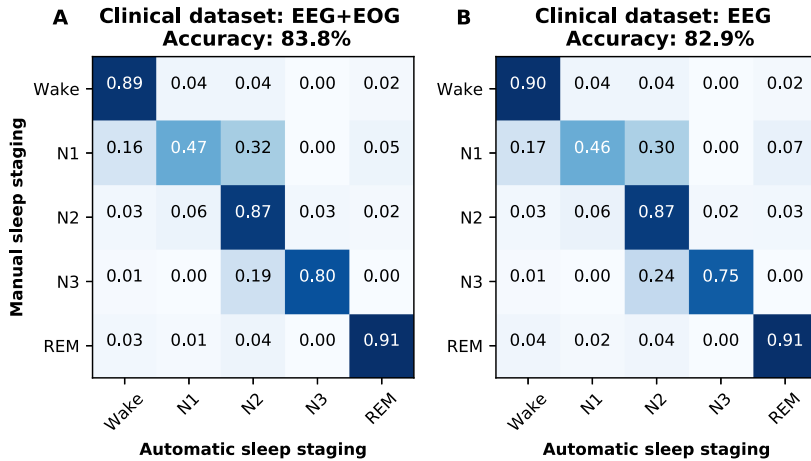


Fig. 3. Normalized confusion matrices of the classification accuracies from the clinical dataset with (A) two-channel input (F4-M1 EEG and E1-M2 EOG) and (B) single EEG channel (F4-M1) input.

sets, respectively. These accuracies corresponded to kappa values of 0.82, 0.78, and 0.77. In the test set, the sensitivity for identifying sleep was 95.6% with 89.8% specificity. The N1 sleep stage was the most challenging to identify (classification accuracy of 46.0%). In contrast, wake was identified with 89.8% accuracy, N2 with 86.5%, N3 with 75.4%, and REM with 90.8% accuracy (Fig. 3B).

C. OSA Severity

When comparing the OSA severity groups, the accuracies and kappa values were lowest for patients with severe OSA (Table IV). The accuracy increased with decreasing OSA severity and were the highest for individuals without OSA. Similar behavior was perceived in the individual sleep stages, with the

exception of N1 sleep which was most accurately classified for severe OSA patients (Fig. 4).

IV. DISCUSSION

In this study, we developed a deep learning-based method for automatic classification of sleep stages from raw EEG and EOG signals using both a large clinical dataset ($n = 891$) comprising patients with suspected OSA and a publicly available dataset of healthy individuals ($n = 153$). Sleep staging was implemented using both two-channel input and single-channel input. Furthermore, we also studied the effect of OSA severity on the performance of automatic sleep staging. Overall, the automatic sleep staging method achieved high accuracies: 83.9% ($\kappa = 0.78$) and 83.6% ($\kappa = 0.77$) with single and two-channel inputs,

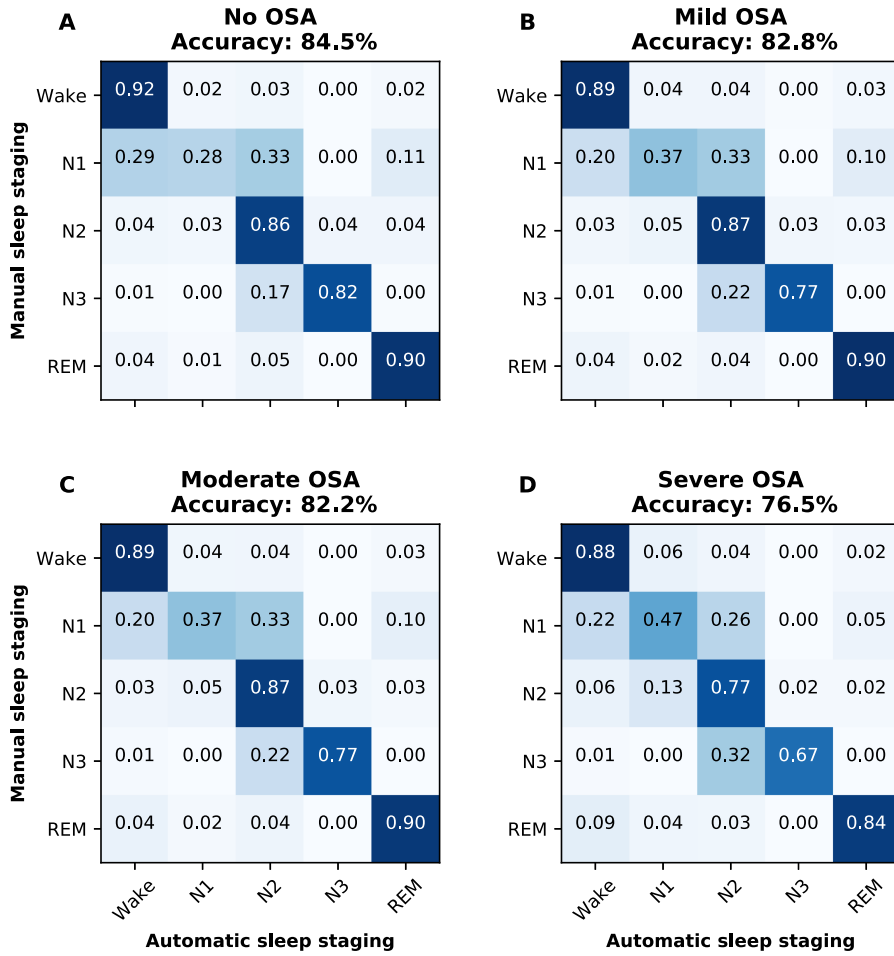


Fig. 4. Normalized confusion matrices of the classification accuracies with single EEG channel (F4-M1) in individuals (A) with no OSA, (B) with mild OSA, (C) with moderate OSA, and (D) with severe OSA.

respectively, in the public dataset, and almost correspondingly 82.9% ($\kappa = 0.78$) and 83.8% ($\kappa = 0.77$) in the clinical dataset. The accuracy of the sleep staging decreased with increasing OSA severity with the accuracy being the highest for individuals without OSA and lowest with individuals having severe OSA. Based on the obtained results, deep learning could enable accurate sleep staging with a single easily measurable frontal EEG channel with practically the same accuracy as with the additional EOG channel. Overall, the reliability of these automatic sleep staging approaches was comparable with the reliability of manual sleep scoring [4]–[9].

The developed deep learning model compared favorably to previous studies based on the publicly available Sleep-EDF dataset [32], [33]. Our method slightly surpassed the performance of previously published methods (Table III). Previously, Mousavi *et al.* have utilized the updated Sleep-EDF dataset with 153 recordings and included only 30 minutes of

TABLE III
PERFORMANCE COMPARISON

	Recordings (n)	Cross-validation	Accuracy	κ
<i>Two-channel: Fpz-Cz and EOG</i>				
Present work	153	10-fold	83.9%	0.78
Phan et al. [19]	39	20-fold	82.3%	0.75
Andreotti et al. [17]	38	20-fold	76.8%	0.68
<i>Single-channel: Fpz-Cz</i>				
Present work	153	10-fold	83.7%	0.77
Mousavi et al. [18]	153	10-fold	80.03%	0.73
Mousavi et al. [18]	39	20-fold	84.26%	0.79
Supratak et al. [11]	39	20-fold	82.0%	0.76
Phan et al. [19]	39	20-fold	81.9%	0.74
Tsinalis et al. [20]	39	20-fold	78.9%	-
Tsinalis et al. [21]	39	20-fold	74.8%	-

Only studies utilizing the sleep cassette dataset of the Sleep-EDF, conducting cross-validation with an independent test set, and having truncated the excess wake periods from the recordings are included.

TABLE IV
PERFORMANCE OF THE DEEP LEARNING-BASED SLEEP STAGING IN OBSTRUCTIVE SLEEP APNEA SEVERITY GROUPS

	<i>n</i>	Accuracy			κ		
		Training	Validation	Test	Training	Validation	Test
No OSA	152	89.4%	84.4%	84.5%	0.86	0.79	0.79
Mild OSA	278	87.7%	82.4%	82.8%	0.83	0.77	0.77
Moderate OSA	208	87.2%	83.0%	82.2%	0.83	0.77	0.76
Severe OSA	254	82.9%	76.7%	76.5%	0.77	0.68	0.68

wake before and after sleep achieving an accuracy of 80.03% ($\kappa = 0.73$) with a single EEG channel [18]. In comparison, we achieved a single-channel accuracy of 83.7% ($\kappa = 0.77$) with the same dataset and identically truncated signals. Other studies based on state-of-the-art methods have been conducted with the smaller Sleep-EDF dataset with only 39 recordings [11], [17], [19]–[21] and thus direct comparison is difficult. However, it is noteworthy that Mousavi *et al.* compared the performance of their sleep staging method in both the smaller and updated datasets and achieved significantly higher accuracy (84.26% vs. 80.03%) in the smaller dataset [18]. This indicates that accurate sleep staging may be easier in the smaller dataset when compared to the larger, updated dataset used in the present study. Furthermore, direct comparison with previous studies is difficult due to non-standardized use of the database. The recordings in the database contain excessive wake periods before and after sleep. Inclusion of the excess wake periods to the automatic sleep staging can lead to overly optimistic results. Therefore, we only compared our results to studies truncating the excess amount of wake either by using only 30 minutes of wake before and after sleep [11], [18] or by only using the sleep [19], [20], [21]. Furthermore, the results cannot be compared to studies not using an independent test set to assess the performance, as these results could be distorted by overfitting.

The PSGs collected from suspected OSA patients have been problematic for previous automatic sleep staging approaches and even the reliability of manual scoring is known to be lower than with healthy individuals [8], [9], [28]. This is most likely due to a fragmented sleep structure and an increase in the amount of N1 sleep stage, which are typical for OSA patients [9]. In the present study, the sleep staging accuracies decreased with increasing OSA severity, with an accuracy of 84.5% for individuals without OSA and 76.5% for patients with severe OSA. Wake and N1 sleep comprised a larger portion of the recording whereas N2, N3, and REM comprised a smaller portion of the recording for patients with severe OSA when compared to the other patient groups (Table II). Especially N1 comprised a significantly larger portion (15%) of the recordings in the severe OSA group compared to the other groups (6–9%). This supports the idea that fragmented sleep structure caused by OSA impairs the accuracy and reliability of sleep staging. However, it is noteworthy that the accuracy of scoring N1 was 47% for patients with severe OSA (Fig. 4D) while it was only 28% for individuals without OSA (Fig. 4A). This increase in accuracy is likely due to a larger amount of N1 sleep epochs and transitions between wake and N1 available during the training of the deep learning model. Furthermore, it is possible that manually identifying the N1

sleep of an individual patient becomes more reliable when more N1 sleep and especially more transitions between wake and N1 are available. This could improve the automatic scoring of N1 in addition to the accuracy increasing simply due to the larger training material. However, the N1 accuracy remained the lowest amongst all sleep stages and the accuracy of the other stages decreased for severe OSA. Thus, the increase in N1 accuracy was insufficient to compensate for the reduction in total accuracy with increasing OSA severity.

Implementation of automatic sleep staging system in a clinical setting could provide significant benefits over the prevailing practice. Currently, the manual sleep staging lacks sufficient inter-rater reliability, as perceived from numerous studies [4]–[9]. It could be argued that since our deep learning-based sleep staging method was trained with manual scorings, its accuracy cannot surpass human scorers. However, the developed automatic method may produce a consensus over multiple scorers and thus minimize the variability. The developed automatic sleep staging method did not learn only from a single scorer as the clinical PSGs were scored by multiple sleep technicians potentially differing in their scoring preferences and traditions. Thus, the optimal solution is not to mimic a single scorer but rather classify the stages as similarly as possible to the majority of the scorers. Furthermore, after training, the automatic method always scores the sleep stages similarly regardless of the situation. This can be a major advantage over a manual scorer, as the automatic scoring does not depend on factors such as human error, vigilance level, or the current scoring environment.

In addition to high variability, manual sleep staging is highly time-consuming and requires trained specialists for a rather repetitive task. The sleep staging of a single patient could be performed in less than a second with the proposed automatic sleep staging method, whereas the manual scoring can take up to hours even for experienced scorers. Although the automatic sleep staging method is reliable for suspected OSA patients, the reliability of sleep stage classification of individuals with other sleep disorders remains to be studied.

Accurate sleep staging with a single EEG channel may present opportunities for further development and application of various ambulatory EEG and PSG acquisition systems [40], [41]. Currently, conducting PSG is expensive and requires trained specialists. Thus, cheaper ambulatory recordings have been developed and shown to be accurate for the diagnosis of OSA [3]. Ambulatory recordings are even the preferred diagnostic method in some healthcare systems [42], [43]. However, the major disadvantage of ambulatory recordings is often the lack of EEG recording, preventing identification of sleep stages and

resulting in crude approximations of the total sleep time from other signals. Thus, ambulatory EEG recording based on a single frontal channel could enhance the accuracy of the ambulatory recordings whilst ensuring simplicity and cost-efficiency. However, further studies are warranted to assess and verify the performance of the developed sleep staging method when applied together with an ambulatory recording device.

The most significant limitation of the developed deep learning-based sleep staging method is the scoring of N1 sleep stage. With both the two-channel and single-channel approaches, the agreement with the manual scoring of stage N1 was the lowest of all sleep stages with a variation of 28–47% between the public and clinical datasets and depending on the severity of OSA. However, N1 is the most difficult sleep stage to identify even for experienced manual scorers [7], [8]. The agreement in N1 we achieved with the automatic sleep staging method is, however, comparable to the inter-rater agreement between manual scorers, which is between 0.19 and 0.46 [4]–[6]. Thus, the limited accuracy of scoring N1 sleep stage may not be due to the developed sleep staging method, but rather in the scoring definitions of N1 resulting in disagreement between experienced manual scorers.

VC ONCLUSION

The proposed deep learning-based automatic method enables reliable, fast, and accurate sleep staging for suspected OSA patients. The accuracy of the sleep staging decreases with increasing OSA severity but with the utilized large clinical dataset, the sleep staging can be conducted for patients suffering from OSA with almost comparable accuracy to individuals without OSA. Practically, automatic sleep staging can be performed as accurately using either a combination of single EEG and EOG signals or using a single frontal EEG channel. The single-channel approach could enable a cost-efficient, simple, and accurate sleep staging in OSA diagnostics.

REFERENCES

- [1] 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, 2017.
- [2] R. B. Berry *et al.*, "The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications," Version 2.5, American Academy of Sleep Medicine, Darien, IL, USA, 2018.
- [3] N. Collop, "Portable monitoring for the diagnosis of obstructive sleep apnea," *Current Opinion Pulmonary Med.*, vol. 14, no. 6, pp. 525–529, 2008.
- [4] 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, 2009.
- [5] U. J. Magalang *et al.*, "Agreement in the scoring of respiratory events and sleep among international sleep centers," *Sleep*, vol. 36, no. 4, pp. 591–596, 2013.
- [6] X. Zhang *et al.*, "Process and outcome for international reliability in sleep scoring," *Sleep Breath.*, vol. 19, no. 1, pp. 191–195, 2015.
- [7] R. S. Rosenberg and S. Van Hout, "The american academy of sleep medicine inter-scorer reliability program: Respiratory events," *J. Clin. Sleep Med.*, vol. 10, no. 4, pp. 447–454, 2014.
- [8] T. Penzel, X. Zhang, and I. Fietze, "Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules," *J. Clin. Sleep Med.*, vol. 9, no. 1, pp. 89–91, 2013.
- [9] 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. 901–908, 2000.
- [10] N. Michielli, U. R. Acharya, and F. Molinari, "Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals," *Comput. Biol. Med.*, vol. 106, pp. 71–81, 2019.
- [11] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1998–2008, Nov. 2017.
- [12] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 4, pp. 758–769, Apr. 2018.
- [13] A. R. Hassan and M. I. H. Bhuiyan, "Automated identification of sleep states from EEG signals by means of ensemble empirical mode decomposition and random under sampling boosting," *Comput. Methods Programs Biomed.*, vol. 140, pp. 201–210, 2017.
- [14] E. Bresch, U. Großekathöfer, and G. Garcia-Molina, "Recurrent deep neural networks for real-time sleep stage classification from single channel EEG," *Frontiers Comput. Neuroscience*, vol. 12, pp. 1–12, 2018.
- [15] A. R. Hassan and M. I. H. Bhuiyan, "An automated method for sleep staging from EEG signals using normal inverse Gaussian parameters and adaptive boosting," *Neurocomputing*, vol. 219, pp. 76–87, 2017.
- [16] A. I. Humayun, A. S. Sushmit, T. Hasan, and M. I. H. Bhuiyan, "End-to-end sleep staging with raw single channel EEG using deep residual convnets," in *Proc. IEEE EMBS Int. Conf. Biomedical Health Inf.*, 2019, pp. 1–5.
- [17] F. Andreotti, H. Phan, N. Cooray, C. Lo, M. T. M. Hu, and M. De Vos, "Multichannel sleep stage classification and transfer learning using convolutional neural networks," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2018, pp. 171–174.
- [18] S. Mousavi, F. Afghah, and U. R. Acharya, "SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PLoS One*, vol. 14, no. 5, 2019, Art. no. e0216456.
- [19] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. De Vos, "Joint classification and prediction CNN framework for automatic sleep stage classification," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1285–1296, May 2019.
- [20] O. Tsinalis, P. M. Matthews, and Y. Guo, "Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders," *Ann. Biomed. Eng.*, vol. 44, no. 5, pp. 1587–1597, 2016.
- [21] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, "Automatic sleep stage scoring with single-channel EEG using convolutional neural networks," 2016, *arXiv preprint arXiv:1610.01683*.
- [22] B. Koley and D. Dey, "An ensemble system for automatic sleep stage classification using single channel EEG signal," *Comput. Biol. Med.*, vol. 42, no. 12, pp. 1186–1195, 2012.
- [23] U. R. Acharya, E. C.-P. Chua, K. C. Chua, L. C. Min, and T. Tamura, "Analysis and automatic identification of sleep stages using higher order spectra," *Int. J. Neural Syst.*, vol. 20, no. 06, pp. 509–521, 2010.
- [24] P. Anderer *et al.*, "An E-health solution for automatic sleep classification according to Rechtschaffen and Kales: Validation study of the somnolyzer 24 × 7 utilizing the siesta database," *Neuropsychobiology*, vol. 51, no. 3, pp. 115–133, 2005.
- [25] S. F. Liang, C. E. Kuo, Y. H. Hu, Y. H. Pan, and Y. H. Wang, "Automatic stage scoring of single-channel sleep EEG by using multiscale entropy and autoregressive models," *IEEE Trans. Instrum. Meas.*, vol. 61, no. 6, pp. 1649–1657, Jun. 2012.
- [26] C. Stepnowsky, D. Levendowski, D. Popovic, I. Ayappa, and D. M. Rapoport, "Scoring accuracy of automated sleep staging from a bipolar electroocular recording compared to manual scoring by multiple raters," *Sleep Med.*, vol. 14, no. 11, pp. 1199–1207, 2013.
- [27] S. Biswal, H. Sun, B. Goparaju, M. Brandon Westover, J. Sun, and M. T. Bianchi, "Expert-level sleep scoring with deep neural networks," *J. Amer. Med. Informat. Assoc.*, vol. 25, no. 12, pp. 1643–1650, 2018.
- [28] A. Patanaik, J. L. Ong, J. J. Gooley, S. Ancoli-Israel, and M. W. L. Chee, "An end-to-end framework for real-time automatic sleep stage classification," *Sleep*, vol. 41, no. 5, pp. 1–11, 2018.
- [29] H. Phan, F. Andreotti, N. Cooray, O. Chén, and M. de Vos, "SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 3, pp. 400–410, Jan. 2019.

- [30] A. Malafeev *et al.*, "Automatic human sleep stage scoring using deep neural networks," *Frontiers Neuroscience*, vol. 12, no. 781, pp. 1–15, 2018.
- [31] H. Sun *et al.*, "Large-Scale automated sleep staging," *Sleep*, vol. 40, no. 10, pp. 1–12, 2017.
- [32] B. Kemp, A. H. Zwirnerman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberyé, "Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 9, pp. 1185–1194, Sep. 2000.
- [33] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, 2000, Art. no. e220.
- [34] A. Rechtschaffen and A. Kales, *A Manual of Standardized Terminology, Techniques and Scoring System of Sleep Stages in Human Subjects*. Los Angeles, CA, USA: Univ. California, 1968.
- [35] AASM, "Sleep-related breathing disorders in adults: Recommendations for syndrome definition and measurement techniques in clinical research," *Sleep*, vol. 22, pp. 667–689, 1999.
- [36] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv preprint arXiv:1608.03983*.
- [37] L. N. Smith, "Cyclical learning rates for training neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vision*, 2017, pp. 464–472.
- [38] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.
- [39] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
- [40] S. Myllymaa *et al.*, "Assessment of the suitability of using a forehead EEG electrode set and chin EMG electrodes for sleep staging in polysomnography," *J. Sleep Res.*, vol. 25, no. 6, pp. 636–645, 2016.
- [41] T. Miettinen *et al.*, "Success rate and technical quality of home polysomnography with self-applicable electrode set in subjects with possible sleep Bruxism," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 4, pp. 1124–1132, Jul. 2018.
- [42] W. W. Flemons, N. J. Douglas, S. T. Kuna, D. O. Rodenstein, and J. Wheatley, "Access to diagnosis and treatment of patients with suspected sleep apnea," *Amer. J. Respiratory Crit. Care Med.*, vol. 169, no. 6, pp. 668–672, 2004.
- [43] E. S. Arnardottir *et al.*, "Variability in recording and scoring of respiratory events during sleep in Europe: A need for uniform standards," *J. Sleep Res.*, vol. 25, no. 2, pp. 144–157, 2016.

Paper III

Korkalainen H, Aakko J, Duce B, Kainulainen S, Leino A, Nikkonen S, Afara IO, Myllymaa S, Töyräs J, and Leppänen T.

Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea

Sleep,

zsaa098 (Advance online publication), 2020.

DOI: 10.1093/sleep/zsaa098

Reprinted under the terms of the Creative Commons Attribution Non-Commercial License (CC BY-NC 4.0)



ORIGINAL ARTICLE

Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea

Henri Korkalainen^{1,2,†,*}, Juhani Aakko^{3,†}, Brett Duce^{4,5}, Samu Kainulainen^{1,2}, Akseli Leino^{1,2}, Sami Nikkonen^{1,2}, Isaac O. Afara^{1,6}, Sami Myllymaa^{1,2}, Juha Töyräs^{1,2,6} and Timo Leppänen^{1,2}

¹Department of Applied Physics, University of Eastern Finland, Kuopio, Finland, ²Diagnostic Imaging Center, Kuopio University Hospital, Kuopio, Finland, ³CGI Suomi Oy, Helsinki, Finland, ⁴Department of Respiratory and Sleep Medicine, Sleep Disorders Centre, Princess Alexandra Hospital, Brisbane, Queensland, Australia, ⁵Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, Australia and ⁶School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, Queensland, Australia

*Corresponding author. Henri Korkalainen, Department of Applied Physics, University of Eastern Finland, Yliopistonranta 1, P.O. BOX 1627, FI-70211 Kuopio, Finland. E-mail: henri.korkalainen@uef.fi

[†]These authors contribute equally to this study.

Abstract

Study Objectives: Accurate identification of sleep stages is essential in the diagnosis of sleep disorders (e.g. obstructive sleep apnea [OSA]) but relies on labor-intensive electroencephalogram (EEG)-based manual scoring. Furthermore, long-term assessment of sleep relies on actigraphy differentiating only between wake and sleep periods without identifying specific sleep stages and having low reliability in identifying wake periods after sleep onset. To address these issues, we aimed to develop an automatic method for identifying the sleep stages from the photoplethysmogram (PPG) signal obtained with a simple finger pulse oximeter.

Methods: PPG signals from the diagnostic polysomnographies of suspected OSA patients ($n = 894$) were utilized to develop a combined convolutional and recurrent neural network. The deep learning model was trained individually for three-stage (wake/NREM/REM), four-stage (wake/N1+N2/N3/REM), and five-stage (wake/N1/N2/N3/REM) classification of sleep.

Results: The three-stage model achieved an epoch-by-epoch accuracy of 80.1% with Cohen's κ of 0.65. The four- and five-stage models achieved 68.5% ($\kappa = 0.54$), and 64.1% ($\kappa = 0.51$) accuracies, respectively. With the five-stage model, the total sleep time was underestimated with a mean bias error (SD) of 7.5 (55.2) minutes.

Conclusion: The PPG-based deep learning model enabled accurate estimation of sleep time and differentiation between sleep stages with a moderate agreement to manual EEG-based scoring. As PPG is already included in ambulatory polygraphic recordings, applying the PPG-based sleep staging could improve their diagnostic value by enabling simple, low-cost, and reliable monitoring of sleep and help assess otherwise overlooked conditions such as REM-related OSA.

Statement of Significance

Sleep staging is the cornerstone of diagnosing sleep disorders. However, the diagnosis of obstructive sleep apnea is increasingly reliant on home-based recordings without the ability for sleep staging due to the lack of EEG recording. This hinders the ability to assess sleep architecture, with total sleep time having to be manually estimated from other signals. This leads to large errors in diagnostic parameters that rely on the accurate determination of sleep time. We developed a novel, deep learning-based sleep staging method relying only on photoplethysmogram measured with a finger pulse oximeter. The deep learning approach enables differentiation of sleep stages and accurate estimation of total sleep time. This could easily enhance the diagnostic yield of home-based recordings and enable cost-efficient, long-term monitoring of sleep.

Key words: deep learning; photoplethysmogram; obstructive sleep apnea; recurrent neural networks; sleep staging

Submitted: 10 December, 2019; Revised: 5 March, 2020

© Sleep Research Society 2020. Published by Oxford University Press [on behalf of the Sleep Research Society].

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Introduction

Characterization of sleep architecture via sleep staging is imperative in the diagnosis of various sleep disorders. Currently, assessment of sleep and its quality is also being integrated into an increasing number of consumer-grade health technology devices developed mainly for self-monitoring purposes. In sleep staging, the night is divided into 30-second periods, i.e. epochs, and a sleep stage is assigned to every epoch: wakefulness, light sleep (stages N1 and N2), deep sleep (stage N3), and rapid eye movement (REM) sleep [1]. These sleep stages are identified by visually inspecting electroencephalogram (EEG), electrooculogram (EOG), and submental electromyogram (EMG) signals. These bioelectric signals are usually recorded during polysomnography (PSG) in addition to cardiorespiratory signals such as respiratory airflow, cardiac activity via electrocardiography (ECG), and blood oxygen saturation via photoplethysmogram (PPG) obtained with a pulse oximeter.

Conducting an in-lab PSG is expensive, requiring the time and effort of multiple trained professionals. PSG also has a negative impact on sleep quality as the patient is forced to sleep in an unfamiliar environment with multiple electrodes and sensors attached [2]. This results in worse sleep efficiency, shorter sleep duration, and longer sleep latency during an in-lab PSG compared with home-based measurements [2, 3]. However, home-based measurements do not usually incorporate a recording of EEG. To overcome these limitations, simple ambulatory EEG recording devices with good recording quality have been introduced [4, 5]. However, despite these recent advances in ambulatory EEG measurement, actigraphy is still the preferred method for assessment of sleep over multiple nights due to its simplicity and low costs [6–8]. Actigraphy relies on sensitive wrist-worn accelerometers (motion sensors) and estimates sleep and wake periods during the night [8]. However, actigraphy tends to overestimate sleep time [8, 9] and is unable to differentiate between sleep stages. Therefore, new simple and cost-effective ambulatory methods and algorithms capable of accurately estimating sleep stages with minimal disruption to sleep are urgently needed.

With recent advances in machine learning, specifically deep learning techniques, automatic sleep staging based on EEG has been successfully demonstrated [10–15]. The EEG recording, however, requires multiple electrodes with meticulous placement. Besides changes in the electrical activity of the brain, sleep stages are reflected in the autonomic nervous system activity. Parasympathetic tone increases when progressing from wake to deep sleep [16, 17], while REM sleep is characterized by increased sympathetic tone [18]. Meanwhile, the sympathetic and parasympathetic tone of wake periods during the night is between those of NREM and REM sleep [19]. It has also been shown that heart rate variability (HRV) differs between sleep stages [16] and that sleep staging with a simpler measurement setup using ECG has the potential to differentiate between wake, light sleep, deep sleep, and REM sleep [20–22]. The ECG-based approaches have relied on HRV features [20] and are often combined with respiratory effort [21] or movement features [22]. Besides ECG, HRV features can be estimated from information contained in the PPG signal [23, 24] recorded during most polygraphic and polysomnographic recordings. Thus, PPG may provide a simpler solution for differentiating between sleep stages.

PPG can be measured with a simple finger pulse oximeter by measuring variations in the transmissive absorption of light related to arterial pulsations. Furthermore, a PPG recording based on reflective absorption is included in many consumer-grade health technology devices such as smartwatches. Recently, there have been attempts to conduct sleep staging using estimated HRV features derived from PPG [25–29]. However, these have focused only on estimating features typically calculated from ECG and have relied on a simultaneous actigraphy recording. However, changes in PPG have also been linked to increased EEG power density and cortical activity during sleep [30] and can be used to determine sympathetic activation [30, 31]. As PPG is related to various physiological characteristics and autonomic nervous system activity, we hypothesize that utilization of deep learning methodology to analyze PPG signal without any prior feature selection enables fast, easily accessible, and accurate sleep staging.

The primary aim of this study was to develop an automatic, deep learning-based sleep staging method utilizing only the PPG signal measured with a transmissive finger pulse oximeter during a full PSG. A secondary aim was to achieve this in an end-to-end manner without any manual feature extraction, i.e. by using the complete PPG signals as recorded with the pulse oximeter and providing the sleep stages automatically for each 30-second segment of the signal. Moreover, we demonstrate the performance of this deep learning approach with three-stage (wake/NREM/REM), four-stage (wake/light sleep (N1+N2)/deep sleep (N3)/REM), and five-stage (wake/N1/N2/N3/REM) classification of sleep and its ability to derive commonly used sleep parameters (total sleep time and sleep efficiency) in a large ($n = 894$) clinical population of patients suspected with obstructive sleep apnea (OSA).

Materials and Methods

Data set

The data set used in this study comprised 933 diagnostic full PSGs conducted due to clinical suspicion of OSA at the Princess Alexandra Hospital (Brisbane, Australia) using Compumedics Grael acquisition system (Compumedics, Abbotsford, Australia) between 2015 and 2017. Approval for data collection was obtained from the Institutional Human Research Ethics Committee of the Princess Alexandra Hospital (HREC/16/QPAH/021 and LNR/2019/QMS/54313). Complete recordings and successful sleep scorings were obtained for 894 patients, yielding the final data set used in this study (Table 1).

Sleep stages were initially scored manually by experienced scorers participating regularly in intra- and interlaboratory scoring concordance activities. A total of 10 scorers participated in the scoring of the whole data set, and each recording was scored once by a single scorer. In a previous study on the interrater reliability at the Princess Alexandra Hospital, the mean (SEM) Cohen's κ of sleep staging was 0.74 (0.02) [32]. As for the individual sleep stages, the κ -values were 0.88 (0.03) for wake, 0.47 (0.08) for N1, 0.68 (0.03) for N2, 0.60 (0.08) for N3, and 0.92 (0.01) for REM [32]. The manual sleep staging was conducted based on EEG, EOG, and chin EMG signals. The sleep stages, arousals, and respiratory events were scored in compliance with the prelaunt American Academy of Sleep Medicine (AASM) guidelines [1].

Table 1. Demographic and polysomnographic information of the study population

	Whole population (n = 894)	Training set (n = 715)	Validation set (n = 90)	Test set (n = 89)
	Median (interquartile range)			
Age (years)	55.9 (44.7–65.8)	55.8 (44.7–66.0)	56.6 (42.9–66.4)	56.1 (45.3–63.3)
Ari (1/h)	20.7 (13.9–31.4)	21.1 (14.1–32.5)	18.9 (13.2–26.6)	20.5 (13.6–29.5)
AHI (1/h)	15.8 (7.0–32.6)	16.0 (7.4–33.5)	12.3 (5.7–30.2)	16.8 (6.5–33.2)
BMI (kg/m ²)	34.4 (29.4–40.4)	34.2 (29.3–40.1)	35.9 (28.6–41.5)	34.8 (31.1–41.2)
N1 (%)	10.9 (6.7–18.8)	11.1 (6.9–19.3)	10.8 (6.0–19.1)	9.7 (5.5–16.2)
N2 (%)	48.3 (41.2–56.2)	48.2 (41.6–56.5)	50.3 (40.3–55.2)	48.8 (38.5–55.6)
N3 (%)	18.3 (9.6–26.8)	18.0 (9.4–26.9)	17.7 (9.4–26.0)	20.4 (11.4–27.8)
NREM (%)	82.9 (77.8–88.1)	83.0 (77.8–88.1)	82.4 (78.5–88.8)	82.4 (77.1–86.4)
REM (%)	17.1 (11.8–22.0)	16.9 (11.8–22.2)	17.5 (11.0–21.4)	17.6 (12.5–22.8)
SE (%)	70.7 (58.1–81.9)	70.7 (57.9–81.7)	69.9 (55.2–83.6)	71.9 (60.1–80.7)
SL (min)	17.5 (9.0–34.5)	17.5 (9.5–35.1)	19.0 (7.0–29.8)	15.0 (9.0–33.5)
TRT (min)	442.3 (409.5–474.0)	442.0 (410.3–474.5)	449.0 (412.4–474.6)	438.0 (403.1–464.5)
TST (min)	308.8 (253.5–359.5)	309.5 (253.0–359.5)	304.0 (249.5–368.6)	304.0 (259.3–347.8)
WASO (min)	102.5 (61.0–149.5)	102.8 (61.0–152.0)	96.0 (60.6–144.4)	100.0 (65.4–135.8)
	n (% of the population)			
No OSA	154 (17.2)	117 (16.4)	20 (22.2)	18 (20.2)
Mild OSA	278 (31.1)	224 (31.3)	29 (32.2)	24 (27.0)
Moderate OSA	209 (23.4)	168 (23.5)	17 (18.9)	24 (27.0)
Severe OSA	253 (28.3)	206 (28.8)	23 (25.6)	24 (27.0)
Female	398 (44.5)	320 (44.8)	39 (43.3)	39 (43.8)
Male	496 (55.5)	395 (55.2)	50 (55.6)	51 (57.3)

Ari, arousal index; AHI, apnea-hypopnea index; BMI, body mass index; SE, sleep efficiency; SL, sleep latency; TRT, total recording time; TST, total sleep time; WASO, wake after sleep onset. No obstructive sleep apnea (OSA): AHI < 5, mild OSA: 5 ≤ AHI < 15, moderate OSA: 15 ≤ AHI < 30, severe OSA: AHI ≥ 30.

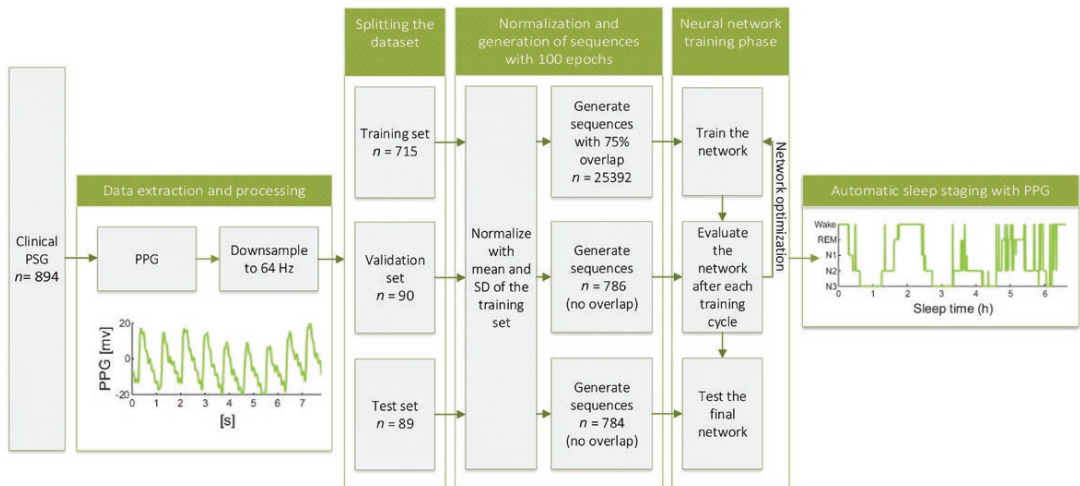


Figure 1. Illustration of the study workflow. The photoplethysmogram (PPG) signals were extracted from clinical polysomnographies (PSG), downsampled, and split into three independent sets: training, validation, and test set. These sets were normalized with z-score normalization using the mean and SD of the training set. The signals were then used to generate sequences of hundred 30-s PPG epochs and an overlap of 75% was used in the training set. The sequences were then used to train, optimize, and test the developed neural network resulting in an automatic sleep staging approach utilizing only PPG signal.

We extracted the transmissive photoplethysmogram (PPG) signals measured with a finger pulse oximeter (Nonin Xpod 3011) from the PSGs with Profusion PSG 4 software (Compumedics, Abbotsford, Australia) and utilized the complete PPG signals without any manual feature selection in the deep learning-based sleep staging. The PPG signals were originally recorded with 256 Hz sampling frequency and were downsampled to 64 Hz in this study to reduce the computational load. No further preprocessing

or any artifact removal was implemented. None of the EEG, EOG, or EMG signals were used beyond the initial manual scoring. The complete study workflow is illustrated in Figure 1.

The complete data set was randomly split into training (715 recordings, 80%), validation (90 recordings, 10%), and test (89 recordings, 10%) sets. Due to the randomization, 85% of the patients in the training set, 78% of the patients in the validation set, and 81% of the patients in the test set had OSA (apnea-hypopnea

index ≥ 5). Subsequently, the data sets were normalized using z-score normalization. To minimize bias, all the data sets were normalized using the mean and SD of the training set. Finally, the PPG signals were divided into 30-second epochs corresponding to the timestamps of the manually scored sleep stages.

Neural network architecture

A convolutional neural network (CNN) combined with a recurrent neural network (RNN) was implemented for sleep stage classification. The classification was conducted individually with three different classification systems: (1) wake, NREM sleep, and REM sleep; (2) wake, light sleep (N1+N2), deep sleep (N3), and REM sleep; and (3) wake, N1, N2, N3, and REM sleep. In essence, CNN was utilized to learn the features of each sleep stage while the RNN was utilized to consider the temporal distribution of sleep stages during the night. The combined CNN and RNN network was implemented in Python 3.6 using Keras API 2.24 with TensorFlow 1.13.1 backend. The implementation of the network is presented in [Supplementary Material](#). The network architecture was identical for the three-, four-, and five-stage classification models.

The CNN consisted of six 1D convolutions, two max-pooling layers, and a global average pooling layer (Figure 2). Each 1D convolution was followed by batch normalization and rectified linear unit activation function. The first 1D convolution had a kernel size of 21 with a stride of 5 and the second 1D convolution had a kernel size of 21 with a stride size of 1. The remaining 1D convolutions had a kernel size of 5 and a stride size of 1. The number of convolutional filters was 64 for the first two convolutions, 128 for the third and fourth convolutions, and 256 for fifth and sixth convolutions. The max-pooling layers were included after the first two convolutions and before the last two convolutions and had a pool size of 2 with a stride size of 2. The last two 1D convolutions were followed by a global average pooling layer.

The RNN included a time distributed layer of the complete CNN described above. The time distributed CNN layer was followed by a gaussian dropout layer with a dropout rate of 0.3 and a bidirectional gated recurrent unit (GRU) layer. The GRU layer comprised 256 cells with a dropout rate of 0.3 in the forward step and 0.5 in the recurrent step. A time distributed dense layer with a softmax activation function was included as the final layer of the model to produce the output sequence of sleep stage probabilities (Figure 2).

The model was trained in an end-to-end manner using sequences of hundred 30-second epochs, and the sleep stages were estimated for each epoch in the sequences. The dimension of a single sequence used as an input to the network was (1, 100, 1920, 1) comprising the number of sequences, length of a sequence (100 epochs in a single sequence), number of data points in a single 30-second epoch with a 64 Hz sampling frequency (1920 data points), and the number of channels (1 PPG channel), respectively. Overlap of 75% between consecutive sequences was applied when forming the sequences in the training set, effectively increasing the size of the training data set fourfold. This procedure was not applied to the validation and test sets. The training set comprised 25 392 sequences, while the validation and test sets comprised 786 and 784 sequences, respectively. The network training was performed using categorical cross-entropy loss function and an Adam optimizer with warm restarts [33] and a learning rate range of 0.001–0.00001. The optimal range for the learning rate was estimated using learning

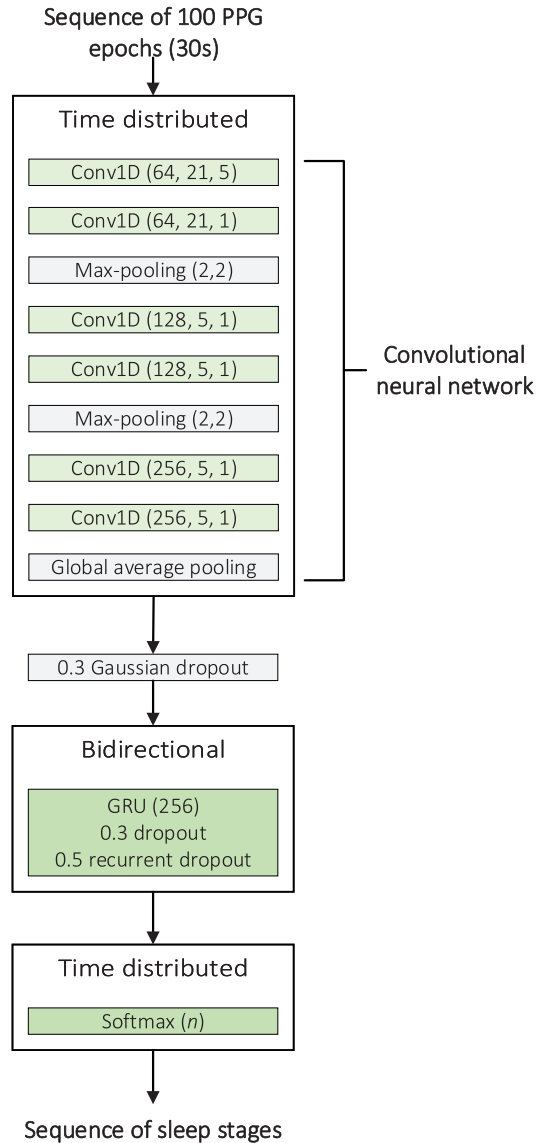


Figure 2. Illustration of the architecture of the combined convolutional neural network (CNN) and recurrent neural network (RNN). The CNN comprised six 1D convolutions (Conv1D), batch normalizations, and rectified linear unit (ReLU) activation functions. The parameters of the convolutional layers are given as (number of filters, kernel size, stride size) and the parameters of the max-pooling layers are given as (pool size, stride size). The CNN was followed by a Gaussian dropout layer, bidirectional gated recurrent unit (GRU), and a time distributed dense layer with a softmax activation function. The dropout rate is given for the dropouts and the number of units is given for the GRU and the final dense layer. n is the number of sleep stages in the classification system and varied between 3, 4, and 5.

rate finder [34]. The model was validated using the validation set after each training cycle, i.e. after the whole training set was used for training the model.

The training was conducted until the validation loss no longer decreased between consecutive training cycles. The

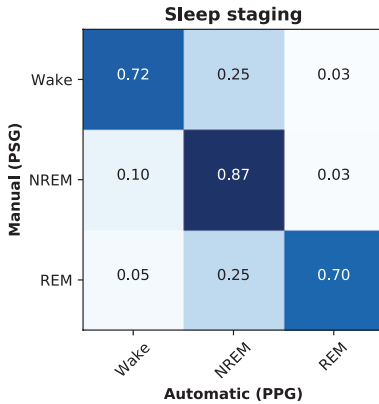


Figure 3. Normalized confusion matrix of the PPG-based classification accuracies for wake, NREM sleep, and REM sleep in an independent test set of 89 patients with suspected obstructive sleep apnea.

model that achieved the lowest validation loss during all the training cycles was considered optimal and was selected for further analysis. The performance of this model was evaluated using the independent test set.

Statistical analysis

The model performance was evaluated by calculating sleep staging accuracies in an epoch-by-epoch manner. Moreover, the inter-rater agreement between the manual PSG-based scoring and automatic PPG-based scoring was assessed using Cohen's kappa coefficient (κ) [35]. Furthermore, the confusion matrices were formed to illustrate the accuracy of each sleep stage and additionally the precision and recall values were calculated.

To further assess the performance of the model, total sleep time, sleep efficiency, and the percentage of sleep stages were calculated from the PPG-based sleep staging and compared with parameters from the manual PSG-based scorings. Furthermore, to study the clinical viability and diagnostic validity of the PPG-based sleep staging, the apnea-hypopnea index (AHI) values derived from the PSGs were compared with those calculated based on the PPG-based sleep staging. When calculating the PPG-AHI, all the respiratory events occurring during epochs scored as wake by the PPG-based sleep staging were discarded and the number of remaining events was divided by the PPG-derived total sleep time. For further comparison, the AHI from polygraphic recordings (PG) was simulated by including all the respiratory events and dividing by the total recording time. The statistical significance of differences was studied using the Wilcoxon signed-rank test in Matlab 2018b (The MathWorks, Natick, MA).

Results

Differentiating between wake, NREM sleep, and REM sleep

In the three-stage classification of sleep (wake/NREM/REM), the deep learning model trained with PPG signals achieved an epoch-by-epoch accuracy of 89.0% in the training set

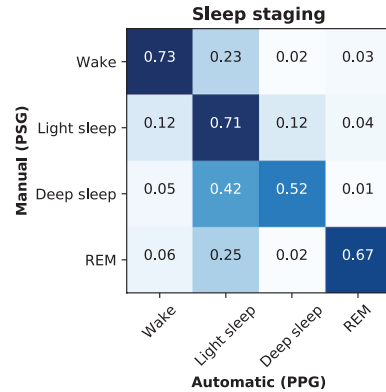


Figure 4. Normalized confusion matrix of the PPG-based classification accuracies for wake, light sleep (N1+N2), deep sleep (N3), and REM sleep in an independent test set of 89 patients with suspected obstructive sleep apnea.

($n = 715$), 79.5% in the validation set ($n = 90$), and 80.1% in the test set ($n = 89$). The accuracies corresponded to Cohen's κ -values of 0.81, 0.63, and 0.65, respectively. For the individual sleep stages in the test set, the precision (recall) was 0.79 (0.72) for wake, 0.81 (0.87) for NREM, and 0.77 (0.70) for REM (Figure 3).

Differentiating between wake, light sleep, deep sleep, and REM sleep

The model developed for the four-stage classification of sleep (wake/N1+N2/N3/REM) achieved an epoch-by-epoch accuracy of 83.1% in the training set, 67.1% in the validation set, and 68.5% in the test set. These corresponded to Cohen's κ -values of 0.75, 0.51, and 0.54 in the training, validation, and test sets, respectively. In the test set, the precision (recall) was 0.78 (0.73) for wake, 0.64 (0.71) for light sleep, 0.57 (0.52) for deep sleep, and 0.75 (0.67) for REM (Figure 4).

Differentiating between wake, N1, N2, N3, and REM sleep

The five-stage (wake/N1/N2/N3/REM) classification model achieved an epoch-by-epoch accuracy of 77.5% in the training set, 62.3% in the validation set, and 64.1% in the test set. The corresponding Cohen's κ -values were 0.69, 0.48, and 0.51. The precision (recall) was 0.74 (0.78) for wake, 0.34 (0.13) for N1, 0.56 (0.67) for N2, 0.61 (0.54) for N3, and 0.75 (0.69) for REM (Figure 5). Examples of the PPG signals during correctly classified sleep stages are presented in Figure 6.

Clinical parameters

Clinical parameters (total sleep time, sleep efficiency, sleep stage percentages, and AHI) were calculated from the manual PSG-based scorings and from the automatic scorings based only on the PPG signal separately for each classification model. In the independent test set, the mean (SD) total sleep time was 298.4 minutes (79.8 minutes) based on the manual scoring. The mean difference to manual scoring was -12.2 minutes (52.9 minutes)

with the three-stage model ($p = 0.03$), -8.8 minutes (55.5 minutes) with the four-stage model ($p = 0.06$), and 7.5 minutes (55.2 minutes) with the five-stage model ($p = 0.24$).

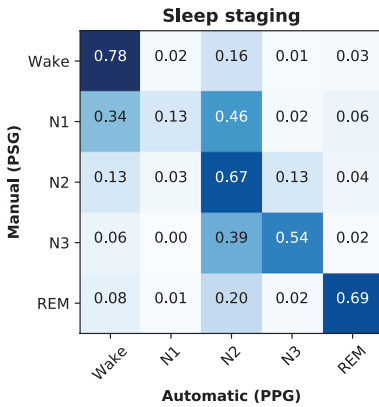


Figure 5. Normalized confusion matrix of the PPG-based classification accuracies for wake, N1, N2, N3, and REM sleep in an independent test set of 89 patients with suspected obstructive sleep apnea.

The mean (SD) sleep efficiency based on the manual scoring was 68.4% (16.9%). The mean difference was -2.8% (11.3%) with the three-stage model ($p = 0.03$), -2.0% (12.0%) with the four-stage model ($p = 0.06$), and 1.9% (12.2%) with the five-stage model ($p = 0.23$). Bland-Altman plots for the total sleep time and sleep efficiency are shown in Figure 7.

The mean (SD) percentage of wake in the test set was 31.6% (16.9%) based on the manual scoring. The difference was 2.7% (11.3%) with the three-stage model ($p = 0.03$), 2.0% (12.0%) with the four-stage model ($p = 0.06$), and -1.9% (12.2%) with the five-stage model ($p = 0.23$). Similarly, the percentage of REM was 12.5% (6.5%) with manual scoring and the differences were 1.1% (5.7%) ($p = 0.05$), 1.3% (5.9%) ($p = 0.08$), and 1.1% (5.7%) ($p = 0.26$) with the three-, four-, and five-stage models, respectively. Percentage of NREM sleep was 55.9% (13.8%) with manual scoring, and the difference was -3.8% (11.9%) ($p = 0.003$) with the three-stage model. Light sleep and deep sleep percentages were 41.2% (13.2%) and 14.7% (11.6%) with manual scoring and the difference was -4.7% (14.1%) ($p = 0.005$) and 1.4% (11.7%) ($p = 0.24$) with the four-stage model, respectively. With the manual scoring, percentages of N1, N2, and N3 were 8.6% (6.6%), 32.7% (10.9%), and 14.7% (11.6%), respectively, and the difference was 5.4% (5.7%) for N1 ($p < 0.001$), -6.3% (12.3%) for N2 ($p < 0.001$), and 1.7% (11.9%) for N3 ($p = 0.08$) with the five-stage model.

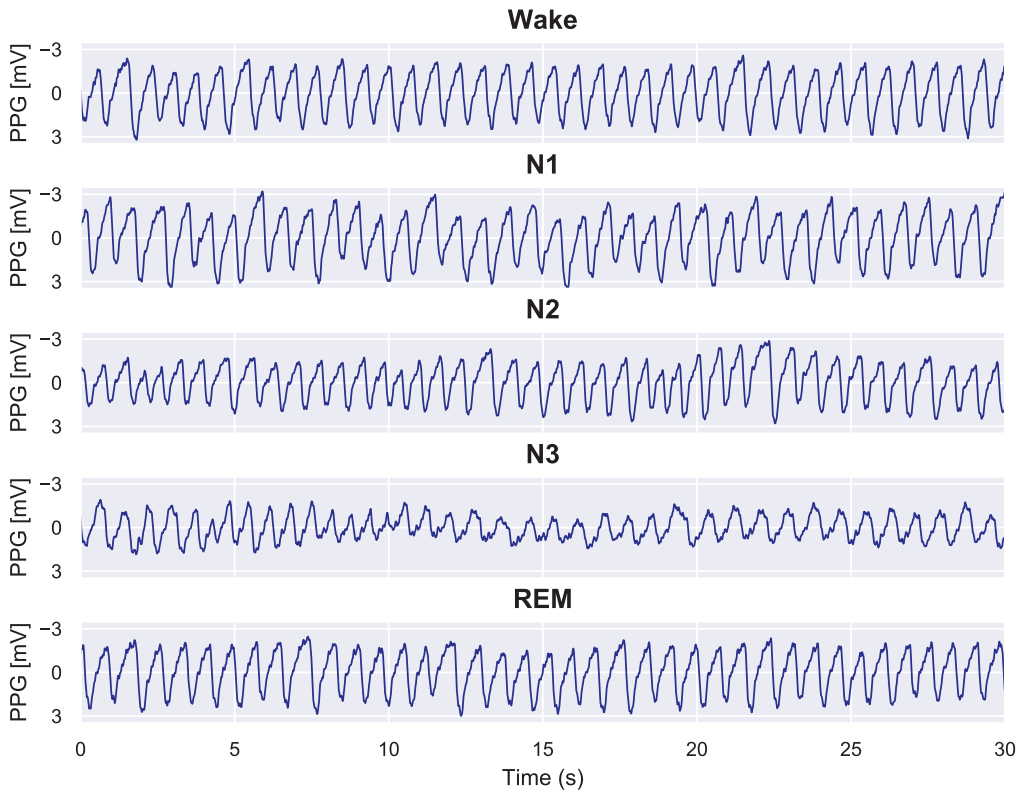


Figure 6. Examples of PPG signals during correctly identified sleep stages. In these examples, it can be seen that during wake the PPG signal remains stable, and the frequency and amplitude are fairly constant. During N1 sleep, irregular variation in the signal amplitude occurs and the frequency decreases. When progressing to N2 and further to N3 sleep, the amplitude decreases and low-frequency oscillations in the PPG signal begin to occur. In contrast, REM sleep is highly similar to wake but with slightly higher variation in the signal amplitude.

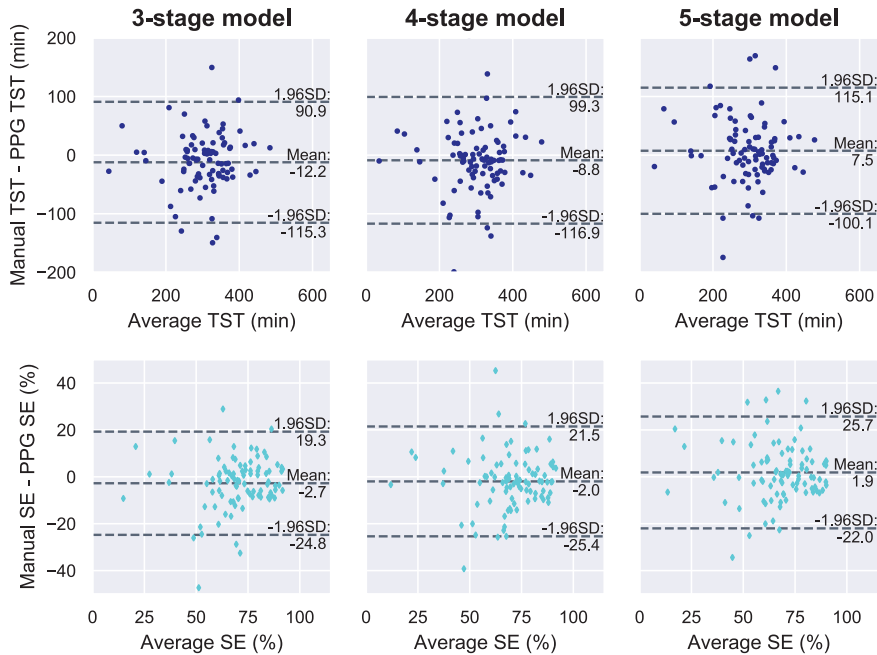


Figure 7. Bland-Altman plots for total sleep time (TST, top row) and sleep efficiency (SE, bottom row) from the deep learning models trained to identify three, four, or five sleep stages. Values are calculated as the average and difference between the values obtained from manual PSG-based sleep scoring and from the automatic PPG-based scoring in an independent test set of 89 patients suspected with obstructive sleep apnea.

The mean (SD) diagnostic AHI calculated from the PSG was 24.2 (24.3) events/h in the test set. The simulated polygraphic AHI was 18.8 (17.5) events/h. With the PPG-based sleep staging, the mean AHI was 23.3 (22.5) events/h with the three-stage model, 23.1 (22.1) events/h with the four-stage model, and 22.6 (22.0) events/h with the five-stage model. The mean difference (SD) between the PSG-AHI and polygraphic AHI was -5.3 (12.4) events/h ($p < 0.001$). The mean difference between the PSG-AHI and PPG-AHI was -0.9 (9.0) events/h with the three-stage model ($p = 0.005$), -1.1 (8.5) events/h with the four-stage model ($p = 0.002$), and -1.6 (8.5) events/h with the five-stage model ($p < 0.001$).

Discussion

In this study, we developed deep learning models for the automated identification of sleep stages from clinical PPG data of suspected OSA patients. The PPG-based sleep staging technique achieved 80.1% epoch-by-epoch accuracy ($\kappa = 0.65$) in three-stage classification (wake/NREM/REM), 68.5% ($\kappa = 0.54$) in four-stage classification (wake/N1+N2/N3/REM), and 64.1% ($\kappa = 0.51$) in five-stage classification (wake/N1/N2/N3/REM) of sleep. Based on the guidelines of Landis and Koch [36], the agreement between manual PSG-based scoring and the developed deep learning-based scoring based solely on PPG was substantial in three-stage classification and moderate in four- and five-stage classification. Therefore, utilization of PPG signal together with deep learning methods appears to be a highly promising approach and may enable sufficiently accurate sleep staging for various applications. For example, in OSA diagnostics, the three-stage classification

might be sufficient to determine the total sleep time and study the disease characteristics in REM or NREM sleep.

In contrast to earlier studies, the present study utilized only the PPG signal in an end-to-end manner producing an easily applicable method for automatic sleep staging. Previous studies have utilized HRV features estimated from PPG signal for sleep staging [25–28]. However, PPG has also been linked to various characteristics generally perceived from EEG. For example, variations in spectral components of EEG during arousals have also been perceived in PPG [30]. This supports using the full PPG signals for the sleep staging instead of just the estimated HRV content.

Previous studies related to PPG-based sleep staging have relied on a relatively small number of healthy individuals (10–152 participants) [25–28] and have often included actigraphy in addition to PPG [25, 26, 28]. In this study, we utilized recordings of 894 individuals with a high prevalence of OSA (83% of the population). Sleep staging of OSA patients is generally more difficult than in healthy population due to fragmented sleep architecture and an increased amount of N1 sleep and sleep stage transitions [37]. Nevertheless, the performance of our algorithm was at least comparable to previous studies. For example, two-stage sleep-wake classification has been previously conducted with 72.36% [29] and 77% accuracy [28], whereas our model achieved an accuracy of 80.1% in three-stage classification (wake/NREM/REM). Similarly, the Cohen's κ -value has been between 0.46 and 0.59 for the three-stage classification [25, 27] and between 0.42 and 0.52 for the four-stage classification (wake/light sleep/deep sleep/REM) [25, 26]. In comparison, we achieved κ -values of 0.65 and 0.54 for the three- and four-stage classification, respectively.

This illustrates that the PPG-based sleep staging could be used beyond healthy individuals and independently without an actigraphy recording.

Accurate sleep monitoring over multiple consecutive nights has been difficult due to the lack of comfortable, wearable sensors that could be used at home without assistance. Actigraphy has been the preferred method for long-term monitoring but is unable to differentiate between sleep stages and overestimates sleep time whenever the individual is awake and motionless in bed [8, 9]. As the PPG recording is comfortable, low cost, and easy to use, the current results suggest that the PPG-based sleep staging could be a reasonable substitute for actigraphy when the ability to differentiate between sleep stages is required.

Application of PPG-based sleep monitoring could improve the information received from ambulatory PG, not including EEG recording. PPG sensors are already integrated into pulse oximeters in ambulatory PG devices; however, in current clinical practice, sleep parameters are qualitatively estimated based on other measured signals, such as movement and breathing. This is possibly the reason for the significant difference in determined sleep time between PG and PSG [38]. For example, in a large European cohort of OSA patients, the mean total sleep time from PSG was 381.7 min, whereas the estimated sleep time from PG was 428.8 minutes [38]. In the present study, the mean bias error (SD) in the estimated total sleep time based on PPG was only 7.5 (55.2) minutes with the five-stage classification. Even though the SD remains relatively large and some outliers in predictions still remain (Figure 7), the PPG-based staging could provide a way to get a sufficiently accurate estimation of total sleep time for most patients. This is an important result since, e.g., in OSA diagnostics the most commonly used diagnostic parameters depend on the total sleep time. For example, the AHI could be determined with a considerably better correspondence to the PSG; the PPG-based AHI differed with only -0.9 events/h from the standard diagnostic AHI whereas, with the simulated PG-AHI the difference was -5.3 events/h.

Furthermore, application of the PPG-based sleep staging and reliable differentiation between wake, NREM, and REM sleep could assist in detecting REM-related OSA from ambulatory PG. When compared with PSG, ambulatory PGs are considerably cheaper to conduct, have better availability, and are already the preferred diagnostic method in some health care systems [39]. Therefore, the application of the PPG-based sleep staging could significantly enhance the already widely used ambulatory PGs and bring their diagnostic value closer to an in-lab PSG without inducing any additional costs. However, further studies are warranted to assess the performance of the PPG-based sleep staging on ambulatory recordings and investigate the effect of common issues related to ambulatory measurements, such as technical problems in data quality, artifacts, and missing sections of the signal during the night. Furthermore, additional studies are warranted to validate the method across different pulse oximeter types and models.

Besides the potential application of PPG-based sleep staging to PG, the method developed in this study could have applications in various consumer-grade health technology devices. Nowadays, reflective PPG sensors are integrated into various wearable self-tracking devices, such as activity wristbands and smartwatches. Such devices already measure sleep duration and quality, but the algorithms implemented in these devices for sleep staging are not public and their validity has not been

thoroughly investigated in a clinical setting [40–43]. In contrast, the PPG-based sleep staging method developed in this study provides highly promising results in a clinical population of patients referred for PSG due to the suspicion of OSA. Thus, it could enable sleep staging beyond the healthy population, enable simple long-term monitoring of sleep quality, and assist in identifying sleep disorders, even with consumer-grade devices. However, the reflective PPG differs from the transmissive measurement giving rise to additional challenges. Therefore, further studies are needed to assess the performance of the developed algorithm in analyzing data from reflective PPG sensors commonly integrated into consumer-grade wearable devices.

The low agreement with manual PSG scoring of N1 is a limitation of the present PPG-based sleep staging. The agreement between manual and PPG-based scoring of the N1 sleep stage was only 13%. The mean percentage of N1 was 8.6% of the recording from the PSG-based scoring while the mean difference was 5.4% with the PPG-based scoring. However, the N1 sleep stage also has a low agreement between manual scorers; the agreement is the lowest of all sleep stages and the κ -value is only between 0.19 and 0.46 [44–46]. This could be the main reason for the low N1 accuracy in the presented PPG-based sleep staging. The N1 sleep stage was mostly misidentified as N2 by the presented PPG-based sleep staging approach. Thus, it is likely that the low N1 agreement is also partially due to relatively small differences in the PPG signals between N1 and N2 sleep stages. This further raises the question of whether differentiating between N1 and N2, the two stages comprising light sleep, is always required for all different applications of sleep staging. Furthermore, the current EEG, EOG, and EMG-based sleep staging suffers from arbitrary rules not fully based on physiological factors. Mainly, the use of 30-second epochs excludes all the information on the sleep microstructure. Therefore, the agreement with PSG-based scoring does not fully capture the feasibility of the PPG sleep staging; rather, future studies are warranted on how the PPG-based sleep staging captures the physiological changes during the night and reflects the outcomes such as perceived sleep quality or daytime vigilance.

In conclusion, as PPG is easy to record, it enables cost-effective and simple sleep monitoring without disrupting natural sleep patterns. Therefore, the PPG-based automatic sleep staging has great potential to supplement the widely used ambulatory PGs, which already include PPG measurement. This could enhance their diagnostic yield by enabling cost-efficient, simple, and reliable long-term monitoring of sleep and by enabling the assessment of otherwise overlooked conditions such as REM-related OSA.

Supplementary Material

Supplementary data are available at SLEEP online.

Funding

This work was financially supported by the Research Committee of the Kuopio University Hospital Catchment Area for the State Research Funding (projects 5041767, 5041768, 5041770, 5041776, 5041779, 5041780, 5041781, and 5041783), the Academy of Finland (decision numbers 313697 and 323536), the Respiratory Foundation of Kuopio Region, the

Research Foundation of the Pulmonary Diseases, Foundation of the Finnish Anti-Tuberculosis Association, the Päivikki and Sakari Sohlberg Foundation, Orion Research Foundation, Instrumentarium Science Foundation, the Finnish Cultural Foundation via the Post Docs in Companies program and via the Central Fund, the Paulo Foundation, the Tampere Tuberculosis Foundation, and Business Finland (decision number 5133/31/2018).

Conflict of interest statement. None declared.

References

- Berry RB, et al. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. Version 2*. Darien, IL: American Academy of Sleep Medicine; 2018. doi:10.1016/j.carbon.2012.07.027
- Bruyneel M, et al. Sleep efficiency during sleep studies: results of a prospective study comparing home-based and in-hospital polysomnography. *J Sleep Res*. 2011;20(1 Pt 2):201–206.
- Iber C, et al. Polysomnography performed in the unattended home versus the attended laboratory setting – sleep heart health study methodology. *Sleep*. 2004;27(3):536–540. doi:10.1093/sleep/27.3.536
- Myllymaa S, et al. Assessment of the suitability of using a forehead EEG electrode set and chin EMG electrodes for sleep staging in polysomnography. *J Sleep Res*. 2016;25(6):636–645.
- Miettinen T, et al. Success rate and technical quality of home polysomnography with self-applicable electrode set in subjects with possible sleep bruxism. *IEEE J Biomed Health Inform*. 2018;22(4):1124–1132.
- Ancoli-Israel S, et al. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*. 2003;26(3):342–392.
- Morgenthaler T, et al.; Standards of Practice Committee; American Academy of Sleep Medicine. Practice parameters for the use of actigraphy in the assessment of sleep and sleep disorders: an update for 2007. *Sleep*. 2007;30(4):519–529.
- Sadeh A. The role and validity of actigraphy in sleep medicine: an update. *Sleep Med Rev*. 2011;15(4):259–267. doi:10.1016/j.smrv.2010.10.001
- Paquet J, et al. Wake detection capacity of actigraphy during sleep. *Sleep*. 2007;30(10):1362–1369.
- Biswal S, et al. Expert-level sleep scoring with deep neural networks. *J Am Med Inform Assoc*. 2018;25(12):1643–1650.
- Patanaik A, et al. An end-to-end framework for real-time automatic sleep stage classification. *Sleep*. 2018;41(5):1–11.
- Phan H, et al. Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Trans Neural Syst Rehabil Eng*. 2019;27(3):400–410.
- Malafeev A, et al. Automatic human sleep stage scoring using deep neural networks. *Front Neurosci*. 2018;12:781.
- Sun H, et al. Large-Scale automated sleep staging. *Sleep*. 2017;40(10). doi:10.1093/sleep/zsx139
- Korkalainen H, et al. Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea. *IEEE J Biomed Heal Informatics*. 2019. doi:10.1109/JBHI.2019.2951346
- Penzel T, et al. Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea. *IEEE Trans Biomed Eng*. 2003;50(10):1143–1151.
- Elsenbruch S, et al. Heart rate variability during waking and sleep in healthy males and females. *Sleep*. 1999;22(8):1067–1071.
- Somers VK, et al. Sympathetic-nerve activity during sleep in normal subjects. *N Engl J Med*. 1993;328(5):303–307.
- Berlad I, et al. Power spectrum analysis and heart rate variability in Stage 4 and REM sleep: evidence for state-specific changes in autonomic dominance. *J Sleep Res*. 1993;2(2):88–90. doi:10.1111/j.1365-2869.1993.tb00067.x
- Li Q, et al. Deep learning in the cross-time frequency domain for sleep staging from a single-lead electrocardiogram. *Physiol Meas*. 2018;39(12):124005.
- Fonseca P, et al. Sleep stage classification with ECG and respiratory effort. *Physiol Meas*. 2015;36(10):2027–2040.
- Willemen T, et al. An evaluation of cardiorespiratory and movement features with respect to sleep-stage classification. *IEEE J Biomed Health Inform*. 2014;18(2):661–669.
- Lu S, et al. Can photoplethysmography variability serve as an alternative approach to obtain heart rate variability information? *J Clin Monit Comput*. 2008;22(1):23–29.
- Selvaraj N, et al. Assessment of heart rate variability derived from finger-tip photoplethysmography as compared to electrocardiography. *J Med Eng Technol*. 2008;32(6):479–484.
- Fonseca P, Weysen T, Goelema MS, et al. Validation of photoplethysmography-based sleep staging compared with polysomnography in healthy middle-aged adults. *Sleep*. 2017;40(7). doi:10.1093/sleep/zsx097
- Beattie Z, et al. Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals. *Physiol Meas*. 2017;38(11):1968–1979.
- Uçar MK, et al. Automatic sleep staging in obstructive sleep apnea patients using photoplethysmography, heart rate variability signal and machine learning techniques. *Neural Comput Appl*. 2018;29(8):1–16.
- Dehkordi P, et al. Sleep/wake classification using cardiorespiratory features extracted from photoplethysmogram. *Comput Cardiol* (2010). 2016;43:1021–1024.
- Motin MA, et al. Sleep-wake classification using statistical features extracted from photoplethysmographic signals. *Conf Proc IEEE Eng Med Biol Soc*. 2019;2019:5564–5567.
- Delessert A, et al. Pulse wave amplitude drops during sleep are reliable surrogate markers of changes in cortical activity. *Sleep*. 2010;33(12):1687–1692.
- Grote L, et al. Finger plethysmography – a method for monitoring finger blood flow during sleep disordered breathing. *Respir Physiol Neurobiol*. 2003;136(2–3):141–152.
- Duce B, et al. The AASM recommended and acceptable EEG montages are comparable for the staging of sleep and scoring of EEG arousals. *J Clin Sleep Med*. 2014;10(7):803–809.
- Loshchilov I, et al. SGDR: stochastic gradient descent with warm restarts. arXiv:1608.03983.
- Smith LN. Cyclical learning rates for training neural networks. In: *Proceeding 2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*; March 24 to 31, 2017; Santa Rosa, USA. New York (NY): IEEE; 2017:464–472.
- Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20(1):37–46.
- Landis JR, et al. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174. doi:10.2307/2529310
- Norman RG, et al. Interobserver agreement among sleep scorers from different centers in a large dataset. *Sleep*. 2000;23(7):901–908.

38. Escourrou P, et al.; ESADA Study Group. The diagnostic method has a strong influence on classification of obstructive sleep apnea. *J Sleep Res.* 2015;**24**(6):730–738.
39. Flemons WW, et al. Access to diagnosis and treatment of patients with suspected sleep apnea. *Am J Respir Crit Care Med.* 2004;**169**(6):668–672.
40. de Zambotti M, et al. The sleep of the ring: comparison of the oura sleep tracker against polysomnography. *Behav Sleep Med.* 2019;**17**(2):124–136.
41. Evenson KR, et al. Systematic review of the validity and reliability of consumer-wearable activity trackers. *Int J Behav Nutr Phys Act.* 2015;**12**:159.
42. Gruwez A, et al. The validity of two commercially-available sleep trackers and actigraphy for assessment of sleep parameters in obstructive sleep apnea patients. *PLoS One.* 2019;**14**(1):e0210569.
43. de Zambotti M, et al. Wearable sleep technology in clinical and research settings. *Med Sci Sports Exerc.* 2019;**51**(7):1538–1557.
44. Danker-Hopfe H, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *J Sleep Res.* 2009;**18**(1):74–84.
45. Magalang UJ, et al.; SAGIC Investigators. Agreement in the scoring of respiratory events and sleep among international sleep centers. *Sleep.* 2013;**36**(4):591–596.
46. Zhang X, et al. Process and outcome for international reliability in sleep scoring. *Sleep Breath.* 2015;**19**(1):191–195.

Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea

Korkalainen H^{1,2}, Aakko J³, Duce B^{4,5}, Kainulainen S^{1,2}, Leino A^{1,2}, Nikkonen S^{1,2}, Afara I O^{1,6}, Myllymaa S^{1,2}, Töyräs J^{1,2,6}, Leppänen T^{1,2}

¹ *Department of Applied Physics, University of Eastern Finland, Kuopio, Finland*

² *Diagnostic Imaging Center, Kuopio University Hospital, Kuopio, Finland*

³ *CGI Suomi Oy, Helsinki, Finland*

⁴ *Sleep Disorders Centre, Princess Alexandra Hospital, Brisbane, Australia*

⁵ *Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Australia*

⁶ *School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia*

1 Supplementary material

In this study, we developed a combined convolutional and recurrent neural network to automatically identify sleep stages from a photoplethysmogram signal obtained with a finger pulse oximeter. The complete neural network structure includes a time distributed layer of the convolutional neural network (CNN) which is then followed by a dropout layer and a bidirectional gated recurrent unit (GRU) layer. The neural network was trained using sequences of hundred 30-second epochs of the PPG signals downsampled to 64 Hz and overlap of 75% was used between consecutive sequences in the training set. The sequence forming protocol was repeated until the complete PPG signals were utilized.

The following sections include the Python implementation of the models using Keras and an example of how the model can be formed and compiled with the functions.

1.1 Python functions of the neural network

Below, the functions used to form the neural network are presented.

```
# The combined convolutional and recurrent neural network developed in the study:  
# H Korkalainen, J Aakko et al. "Deep learning enables sleep staging from  
# photoplethysmogram for patients with suspected sleep apnea"
```

```
# Contact information: henri.korkalainen@uef.fi
```

```
from keras.layers import Input, Conv1D, BatchNormalization
```

```

from keras.layers import Activation, MaxPooling1D, GlobalAveragePooling1D
from keras.layers import Dense, GaussianDropout, TimeDistributed
from keras.layers import Bidirectional, GRU
from keras.models import Model

def build_base_cnn_model(input_size = 1920, input_channels = 1,
                        init_conv_kernel_size = 21, ksize = 5,
                        conv_stride_first = 5,
                        conv_stride_rest = 1,
                        maxpool_size = 2, maxpool_stride = 2,
                        act_function = "relu", fs =64):
    # input_size = length of a single PPG-epoch
    # input_channels = number of PPG channels
    # init_conv_kernel_size = kernel size for the first convolution layer
    # ksize = kernel size for the rest convolutional layers
    # conv_stride_first = stride for the first convolution layer
    # conv_stride_rest = stride for the remaining convolution layers
    # maxpool_size = pool size for the max-pooling
    # maxpool_stride = stride size for the max-pooling
    # act_function = activation function
    # fs = the sampling frequency

    input1 = Input(shape = (input_size, input_channels), name = "input")
    x = Conv1D(fs, kernel_size = init_conv_kernel_size,
              strides = conv_stride_first)(input1)
    x = BatchNormalization()(x)
    x = Activation(act_function)(x)

    x = Conv1D(fs, kernel_size = init_conv_kernel_size,
              strides = conv_stride_rest)(x)
    x = BatchNormalization()(x)
    x = Activation(act_function)(x)

    x = MaxPooling1D(pool_size = maxpool_size, strides = maxpool_stride)(x)

    x = Conv1D(2*fs, kernel_size = ksize, strides = conv_stride_rest)(x)
    x = BatchNormalization()(x)
    x = Activation(act_function)(x)

    x = Conv1D(2*fs, kernel_size = ksize, strides = conv_stride_rest)(x)
    x = BatchNormalization()(x)
    x = Activation(act_function)(x)

    x = MaxPooling1D(pool_size = maxpool_size, strides = maxpool_stride)(x)

    x = Conv1D(4*fs, kernel_size = ksize, strides = conv_stride_rest)(x)
    x = BatchNormalization()(x)
    x = Activation(act_function)(x)

    x = Conv1D(4*fs, kernel_size = ksize, strides = conv_stride_rest)(x)

```

```

x = BatchNormalization()(x)
x = Activation(act_function)(x)

out = GlobalAveragePooling1D()(x)
model = Model(inputs = input1, outputs = out)
return model

def build_cnn_to_rnn_model(input_size = 1920, input_channels = 1, n_categories = 3,
                           seq_length = None, init_conv_kernel_size = 21, ksize = 5,
                           conv_stride_first = 5, conv_stride_rest = 1,
                           maxpool_size = 2, maxpool_stride = 2,
                           gru_units_multiplier = 4, rdo = 0.5, do = 0.3, gdo = 0.3,
                           act_function = "relu", fs = 64):
    # input_size = length of a single PPG-epoch
    # input_channels = number of PPG channels
    # n_categories = number of different sleep stages
    # seq_length = number of epochs in the input.
    # With seq_length = 'None' the model accepts any sequence length
    # init_conv_kernel_size = kernel size for the first convolution layer
    # ksize = kernel size for the rest convolutional layers
    # conv_stride_first = stride for the first convolution layer
    # conv_stride_rest = stride for the remaining convolution layers
    # maxpool_size = pool size for the max-pooling
    # maxpool_stride = stride size for the max-pooling
    # gru_units_multiplier * fs = number of GRU units
    # rdo = recurrent dropout size
    # do = (forward) drop out size
    # gdo = Gaussian dropout size
    # act_function = activation function
    # fs = the sampling frequency

    seq_input = Input(shape=(seq_length, input_size, input_channels))

    base_model = build_base_cnn_model(input_size = input_size,
                                      input_channels = input_channels,
                                      init_conv_kernel_size = init_conv_kernel_size,
                                      ksize = ksize,
                                      conv_stride_first = conv_stride_first,
                                      conv_stride_rest = conv_stride_rest,
                                      maxpool_size = maxpool_size,
                                      maxpool_stride = maxpool_stride,
                                      act_function=act_function,
                                      fs = fs)

    encoded_sequence = TimeDistributed(base_model)(seq_input)

    encoded_sequence = GaussianDropout(gdo)(encoded_sequence)

    encoded_sequence = Bidirectional(GRU(gru_units_multiplier*fs,
                                         return_sequences = True,

```

```
        recurrent_dropout = rdo,
        dropout = do))(encoded_sequence)

out = TimeDistributed(Dense(n_categories, activation = "softmax"))(encoded_sequence)

model = Model(inputs = seq_input, outputs = out)

return model
```

1.2 Compiling the neural network

Below, an example of how the model can be formed and compiled is presented. The input must be a 4D tensor with shape (number of sequences, length of a single sequence, sampling frequency * 30 s, number of channels).

```
cnnrnn = build_cnn_to_rnn_model(input_channels = n_channels,
                               n_categories = n_stages, act_function = "relu",
                               do = 0.3, rdo = 0.5, gdo = 0.3)

cnnrnn.compile(loss = 'categorical_crossentropy', optimizer =
               'adam', metrics = ['accuracy'])
```



HENRI KORKALAINEN

Sleep apnea is a highly common sleep disorder with detrimental effects on general wellbeing and health; however, it often remains undiagnosed as the current clinical practice relies on time-consuming and labor-intensive recordings with limited availability.

This thesis presents novel deep learning-based methods enabling easy and efficient assessment of sleep quality and illustrates how the current diagnostic methods of sleep apnea could be improved.



UNIVERSITY OF
EASTERN FINLAND



uef.fi

**PUBLICATIONS OF
THE UNIVERSITY OF EASTERN FINLAND**
Dissertations in Forestry and Natural Sciences

ISBN 978-952-61-3468-0
ISSN 1798-5668