Deep Learning Methods for Clinical Sleep Analysis

Olesen, Alexander Neergaard

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Olesen, A. N. (2020). *Deep Learning Methods for Clinical Sleep Analysis*. DTU Health Technology.

Deep Learning Methods for Clinical Sleep Analysis

Alexander Neergaard Olesen

PhD Thesis, April 2020

DTU Health Tech
Department of Health Technology

Alexander Neergaard Olesen: *Deep Learning Methods for Clinical Sleep Analysis,* An Exploration in Computational Sleep Science, April 2020

SUPERVISORS:
Main supervisor, Associate Professor MSK, Helge. B. D. Sørensen, PhD, MSc. E.E.
Clinical supervisor, Professor Poul Jennum, MD, PhD
Clinical supervisor, Professor Emmanuel Mignot, MD, PhD

LOCATION:
Kgs. Lyngby

TIME FRAME:
December 2016–April 2020

This thesis was prepared with `classicthesis v4.6` for digital use only and is best viewed on a two-page display.

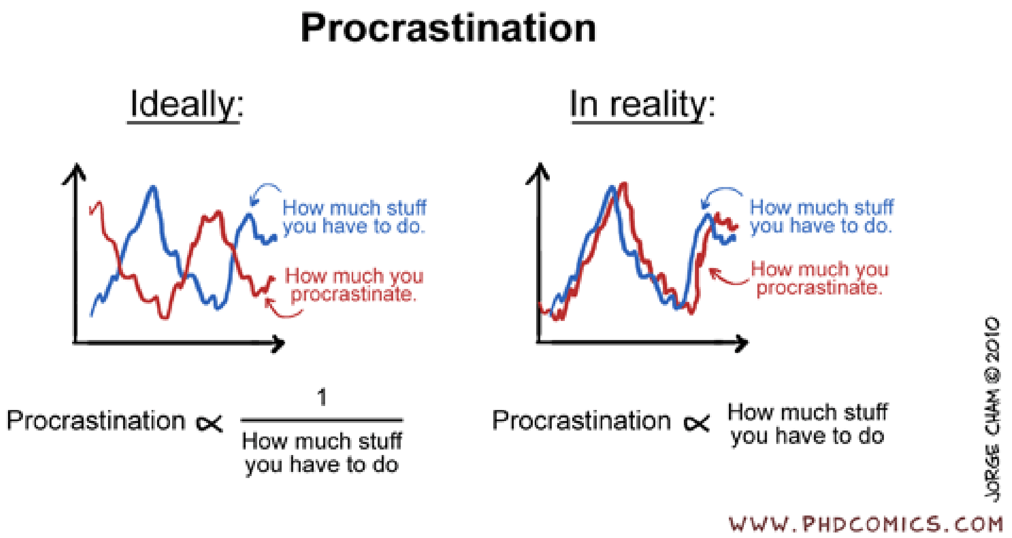
DEEP LEARNING METHODS FOR CLINICAL SLEEP ANALYSIS

ALEXANDER NEERGAARD OLESEN

An Exploration in Computational Sleep Science

PhD Thesis, April 2020
Section for Digital Health
Department of Health Technology
Technical University of Denmark

Supervisors:
Main supervisor, Associate Professor MSK, Helge. B. D. Sørensen, PhD, MSc. E.E.
Clinical supervisor, Professor Poul Jennum, MD, PhD
Clinical supervisor, Professor Emmanuel Mignot, MD, PhD


Procrastination
Ideally:
How much stuff you have to do.
How much you procrastinate.
Procrastination ∝ 1 / How much stuff you have to do
In reality:
How much stuff you have to do.
How much you procrastinate.
Procrastination ∝ How much stuff you have to do
JORGE CHAM © 2010
WWW.PHDCOMICS.COM

CONTENTS

III OUTLOOK

IV APPENDIX

PREFACE

This thesis represents the collective work as a PhD student at the Technical University of Denmark (DTU) as part of the North-Atlantic Research Collaboration between DTU, the Danish Center for Sleep Medicine, and Stanford University, and is submitted as partial fulfillment of the requirements for the degree of Doctor of Philosophy at DTU.

The research presented in this thesis was carried out in the Biomedical Signal Processing Group at the Department of Electrical Engineering, DTU, between December 2016 and September 2017, as a visiting student researcher at Stanford University between October 2017 and June 2019, and in the Section for Digital Health at the Department of Health Technology, DTU, from August 2019 to April 2020 following the formation of the new health department.

Apart from engaging in pure research activities, the PhD studies also presented opportunities to engage as a teaching assistant at DTU, co-supervision of student projects, as well as participating and presenting in international conferences both technical and medical in nature.

The dissertation consists of this summary report, which is based on six research papers and manuscripts written in the period 2016 to 2020. Four of them have been published or accepted for publication, one is currently under review, and the last is currently in preparation.

Kgs. Lyngby, April 2020

Alexander Neergaard Olesen

ACKNOWLEDGMENTS

First of all, I would like to express my deepest gratitude to my main supervisor *Helge B. D. Sørensen*, and clinical co-supervisors *Poul Jennum*, and *Emmanuel Mignot*, for believing in me and allowing me to work on something that I truly enjoy. Thank you for your guidance and support during the last three and a half years.

I would also like to thank all the people associated with the Sleep Lab at Stanford University for welcoming me into their impressive scientific environment, in particular *Eileen Leary*, *Hyatt Moore*, and *Aditya Ambati*; a big thanks to *Stephanie Lettieri* for all your help in general; and also my fellow visiting student researcher *Stanislas Chambon* for engaging and enlightening discussions about machine learning, and for convincing me to switch to PyTorch.

Thank you to all my colleagues at the Section for Health Technology for interesting scientific discussions in our biweekly meetings; my fellow PhD students *Matteo Cesari*, *Mads Olsen*, *Umaer Hanif*, *Andreas Brink-Kjær*, and *Søren Møller Rasmussen* for all of our scientific discussions, and for making sure that I get enough coffee. A particular thanks goes to *Julie A. E. Christensen*. Thank you for your friendship and for always being willing to offer scientific insights and inputs to manuscripts and projects.

A huge thanks goes out to all of the students that I have been engaged with during my time at Stanford: *Lorenzo*, *Andreas*, *Kristina*, *Rasmus*, *Jonathan*, *Jakob*, *Thomas*, *Vicente*, *Aske*, and *Joakim*; many of whom have become good friends, and some even my fellow PhD students. I hope you learned as much from me as I learned from you. Especially, I would like to express my gratitude towards *Jens. B. Stephansen*, who invited me to work with him on the narcolepsy detector, which has been an incredible experience.

I would also like to extend a huge thanks to *Matteo Cesari*, *Martin Nørgaard*, *Sebastian Holst*, *Rasmus Malik Thaarup Høegh*, and *Julie A. E. Christensen*. You have provided valuable comments to this thesis.

To all my friends and family: you have been sorely missed, and I can't wait to hang out and spend time with you all again pending no further restrictions due to corona-virus. A big thank you to my wonderful in-laws, *Annette & Amr*, for allowing me to set up a make-shift office in your dining room and babysitting Theodor during the corona-crisis. To my father *Lars-Ulrik*; thank you for always believing in me, for your guidance and our lunches together at DTU.

And finally, from the bottom of my heart; I want to acknowledge and thank *Alexandra*, the love of my life, for your enthusiasm and willingness to go on an adventure and live in California for almost two years while carrying our first child; for your infinite patience, resilience, love and understanding while I tried my hands at science—without you, this would not have been possible.

ABSTRACT

Sleep disorders are prevalent in the general population and have major implications for personal health and mortality including increased risk of cardiovascular, metabolic and psychiatric complications. Furthermore, sleep disorders have a major economic burden contributing to an estimated cost of several hundred billion dollars per year to society.

The current gold standard for sleep disorder diagnosis is based on manual analysis of electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), and cardio-respiratory variables recorded during sleep following a set of guidelines provided by the American Academy of Sleep Medicine. However, several studies have shown that this introduces subjective bias in the resulting sleep analysis due to interpretations of scoring guidelines, differences in recording equipment and setup among different sleep clinics, and the presence of sleep disorders interrupting normal sleep patterns.

The main objective of this thesis is to develop a system based on artificial intelligence, that can assist clinicians in the analysis of sleep studies. This is realized in three research themes, each focusing on a separate aspect of sleep analysis.

The first research theme presents findings related to the development of methods for automatic sleep stage classification, which is a crucial part of analyzing sleep patterns. Two models based on deep neural networks applied to EEG, EOG, and EMG signals were developed for this reason. Using the raw signals from 14 086 polysomnographies (PSGs) to classify sleep stages, the first model obtained an average accuracy of 86.9% across 1584 PSGs collected from five independent datasets. The second model used cross-correlation representations of signals from 2784 PSGs to classify sleep stages with an accuracy of 86.8% across 70 PSGs scored by six sleep technicians.

The second research theme concerned methods for automatic detection of sleep events focusing specifically on arousals (Ars), limb movements (LMs), and sleep disordered breathing (SDB) events. A model was designed based on deep neural networks applied to EEG, EOG, EMG, and respiratory signals. The model was able to precisely localize and classify events in data and was tested on more than 1000 PSGs. Moreover, the model was used in a transfer learning setting, where a fine-tuning optimization strategy could effectively recover lost performance caused by a reduced set of input channels. An adaptable model like this would be an important step forward in a clinical setting.

The third and final research theme concerned classification of sleep disorders using artificial intelligence. A model was designed based on feature engineering of the hypnodensity-representation and probabilistic classification algorithms to classify narcolepsy type 1 (NT1) from both healthy controls and patients with other central hypersomnias. NT1 was identified with 91% sensitivity and 96% specificity in the test sample, while replication in two independent datasets yielded similar performances.

In conclusion, this thesis presents new automatic methods for clinical sleep analysis based on artificial intelligence. Compared to current methods, the proposed models could significantly reduce analysis time by virtue of being quick to execute, while providing similar or higher levels of performance.

RESUMÉ

Søvnforstyrrelser er udbredte i samfundet og har konsekvenser for personlig sundhed og dødelighed, herunder øget risiko for hjertekarsygdomme, metaboliske og psykiatriske komplikationer. Derudover er søvnforstyrrelser en stor omkostning for samfundet, og det er anslået at komplikationer følgende søvnforstyrrelser bidrager til flere hundrede milliarder dollars om året.

Den nuværende standard for diagnose af søvnforstyrrelse og søvnsygdomme er baseret på manuel analyse af elektroencephalografi (EEG), elektrooculografi (EOG), eletromyografi (EMG) og cardio-respiratoriske variable registreret under søvn efter et sæt retningslinjer defineret af American Academy of Sleep Medicine. Flere studier har imidlertid vist, at dette introducerer et subjektivt bias i den resulterende søvnanalyse på grund af fortolkninger af scoringsretningslinjer, forskelle i optageudstyr og -opsætning blandt forskellige søvnklinikker samt tilstedeværelsen af søvnforstyrrelser, der negativt påvirker normale søvnmønstre.

Hovedformålet med denne afhandling er at udvikle et system baseret på kunstig intelligens, der kan hjælpe klinikere i analysen af søvnmønstre. Dette realiseres i tre forskningstemaer, der hver især fokuserer på et separat aspekt af søvnanalyse.

Først præsenteres resultater relateret til udvikling af metoder til automatisk klassificering af søvnstadier, som er en vigtig del af analysen af søvnmønstre. To modeller baseret på dybe neurale netværk anvendt på EEG, EOG og EMG signaler blev udviklet af denne grund. Ved hjælp af de rå signaler fra 14086 polysomnografier (PSG) til klassificering af søvnstadier opnåede den første model en gennemsnitlig nøjagtighed på 86.9% målt over 1584 PSG'er samlet fra fem uafhængige datasæt. Den anden model brugte krydskorrelations-repræsentationer af signaler fra 2784 PSG'er til klassificering af søvnstadier med en nøjagtighed på 86.8% målt henover 70 PSG'er scoret af seks søvnteknikere.

Dernæst præsenteres en metode til automatisk detektion af mikro-opvågninger, benspjæt, og apnø-lignende perioder. En model blev designet baseret på dybe neurale netværk anvendt på EEG, EOG, EMG og åndedrætssignaler. Modellen var i stand til at lokalisere og klassificere begivenheder i data og blev testet på mere end 1000 PSG'er. Desuden blev modellen brugt til overførselslæring, hvor en finjusteret optimeringsstrategi af modellen effektivt kunne gendanne tabt ydeevne forårsaget af et reduceret sæt inputkanaler. En fleksibel model som denne vil være nyttig i en klinisk sammenhæng.

Det sidste forskningstema vedrørte klassificering af søvnforstyrrelser ved hjælp af kunstig intelligens. En model blev designet baseret på repræsentationer af hypnodensiteten og probabilistiske algoritmer til adskillelse af narkolepsi type 1 (NT1) patienter fra kontroller og patienter med andre centrale hypersomnier. NT1 blev identificeret med 91% sensitivitet og 96% specificitet i testprøven, mens replikation i to uafhængige datasæt gav tilsvarende resultater.

Samlet set præsenterer denne afhandling nye automatiske metoder til klinisk søvnanalyse baseret på kunstig intelligens. Sammenlignet med de nuværende analyse-standarder kan de foreslåede modeller markant reducere analysetiden i kraft af at være hurtigere end manuel analyse, samtidig med at de kan levere lignende eller højere ydeevne.

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AASM	American Academy of Sleep Medicine
AHC	Austrian Hypersomnia Cohort
AHI	Apnea-hypopnea index
AI	Artificial intelligence
ASDA	American Sleep Disorders Association
ANOVA	Analysis of variance
Ar	Arousal
ArI	Arousal index
BF	Basal forebrain
bGRU	Bidirectional gated recurrent unit
BMI	Body-mass index
BN	Batch normalization
CC	Cross-correlation
CNC	Chinese Narcolepsy Cohort
CNN	Convolutional neural network
CSF	Cerebrospinal fluid
DHC	Danish Hypersomnia Cohort
DMH	Dorsomedial hypothalamic nucleus
DRN	Dorsal raphe nucleus
ECG	Electrocardiography
EEG	Electroencephalography
EMG	Electromyography
EOG	Electrooculography
FF	Feed-forward
FHC	French Hypersomnia Cohort
GABA	Gamma-aminobutyric acid
GP	Gaussian process
GRU	Gated recurrent unit
HLA	Human leukocyte antigen
hcrt	Hypocretin
ICC	Intraclass correlation coefficient
ICSD	International Classification of Sleep Disorders
IHC	Italian Hypersomnia Cohort
IIR	Infinite impulse response
IoU	Intersection over union
ISR	Inter-Scorer Reliability program
IS-RC	Interscorer Reliability Cohort
ISRUC	Institute of Systems and Robotics, University of Coimbra Sleep Cohort
JCTS	Jazz Clinical Trial Sample
KHC	Korean Hypersomnia Cohort

LAMF	Low amplitude, mixed frequency
LC	Locus coeruleus
LDT	Laterodorsal tegmental nucleus
LHA	Lateral hypothalamic area
LM	Limb movement
LMI	Limb movement index
LOCI	Leave-one-cohort-in
LOCO	Leave-one-cohort-out
LPT	Lateral pontine tegmentum
LSTM	Long short-term memory
MASSC	Multi-modal automatic sleep stage classification
MCH	Melanin-concentrating hormone
MnPO	Median preoptic nucleus
MrOS	Osteoporotic Fractures in Men Sleep Study
MSED	Multi-modal sleep event detection
MSL	Mean sleep latency
MSLT	Multiple sleep latency test
N1	Non-rapid eye movement stage 1
N2	Non-rapid eye movement stage 2
N3	Non-rapid eye movement stage 3
NREM	Non-rapid eye movement
NSRR	National Sleep Research Resource
NT1	Narcolepsy type 1
NT2	Narcolepsy type 2
OSA	Obstructive sleep apnea
PB	Parabrachial nucleus
PC	Precoeruleus nucleus
PGO	Ponto-geniculo-occipital
PLM	Periodic leg movement
PLMD	Periodic leg movement disorder
PLMI	Periodic leg movement index
POA	Preoptic area
PPT	Pedunculopontine tegmental nucleus
PSG	Polysomnography
RBD	REM sleep behaviour disorder
RDI	Respiratory disturbance index
REM	Rapid eye movement
REML	REM sleep latency
ReLU	Rectified linear unit
RFE	Recursive feature elimination
RNN	Recurrent neural network
ROC	Receiver operating characteristic
SCN	Suprachiasmatic nucleus
SDB	Sleep disordered breathing

SEM	Slow eye movement
SHHS	Sleep Heart Health Study
SLD	Sublaterodorsal nucleus
SOREMP	Sleep onset REM period
SSC	Stanford Sleep Cohort
STAGES	Stanford Technology Analytics and Genomics in Sleep
SWA	Slow wave activity
SWS	Slow wave sleep
TMN	Tuberomammillary nucleus
TST	Total sleep time
vlPAG	Ventrolateral periaqueductal gray
VLPO	Ventrolateral preoptic nucleus
vM	Ventral medulla
vPAG	Ventral periaqueductal gray
WASO	Wake after sleep onset
WSC	Wisconsin Sleep Cohort
W	Wakefulness

Part I

INTRODUCTION

1 THESIS INTRODUCTION

Quantum carburetor? Jesus, Morty, you can't just add a sci-fi word to a car word and hope it means something.

— Rick Sanchez
Rick and Morty, season 2, episode 6.

Although sleep is essential for normal human brain development and functionality, sleep disorders are prevalent in society. There are approximately 90 different sleep disorders currently recognized and described in the International Classification of Sleep Disorders (ICSD) grouped into six categories: insomnias, circadian rhythm sleep-wake disorders, central hypersomnias (e. g. narcolepsy), sleep-related breathing disorders (e. g. obstructive sleep apnea), parasomnias (e. g. sleepwalking, REM sleep behaviour disorder), and sleep-related movement disorders (e. g. periodic leg movement disorder and restless legs syndrome) [1]. It is estimated that about 25% of the US population present with sleep apnea-related symptoms [2], [3] with similar prevalences in other developed countries [4]–[7]. The prevalence of chronic insomnia is estimated to be about 10% of the US population [8], and 6% in high-income countries [9]—a number increasing to up to 48% when including insomnia symptoms alone [9]. Although these numbers are high, evidence suggests that sleep disorders are severely under-diagnosed [10]–[12].

Disrupted sleep is associated with increased risk of developing systemic hypertension, cardiovascular disease and abnormalities in the metabolism [13]. Increased levels of fatigue due to disrupted nighttime sleep is also a cause of motor-vehicle accidents [14], as people with excessive daytime sleepiness have a sevenfold greater risk of being involved in an accident [15].

Apart from the medical impacts on a personal level, sleep disorders have monumental societal impact due to their prevalence and cost of care. A study on the burden of poor sleep in the Australian population estimated the annual economic cost at $42.5 billion [16], while the combined cost of poor sleep in USA, Canada, UK, Japan and Germany is estimated to exceed $600 billion per year [17]. USA alone accounts for an estimated $411 billion of these costs [18].

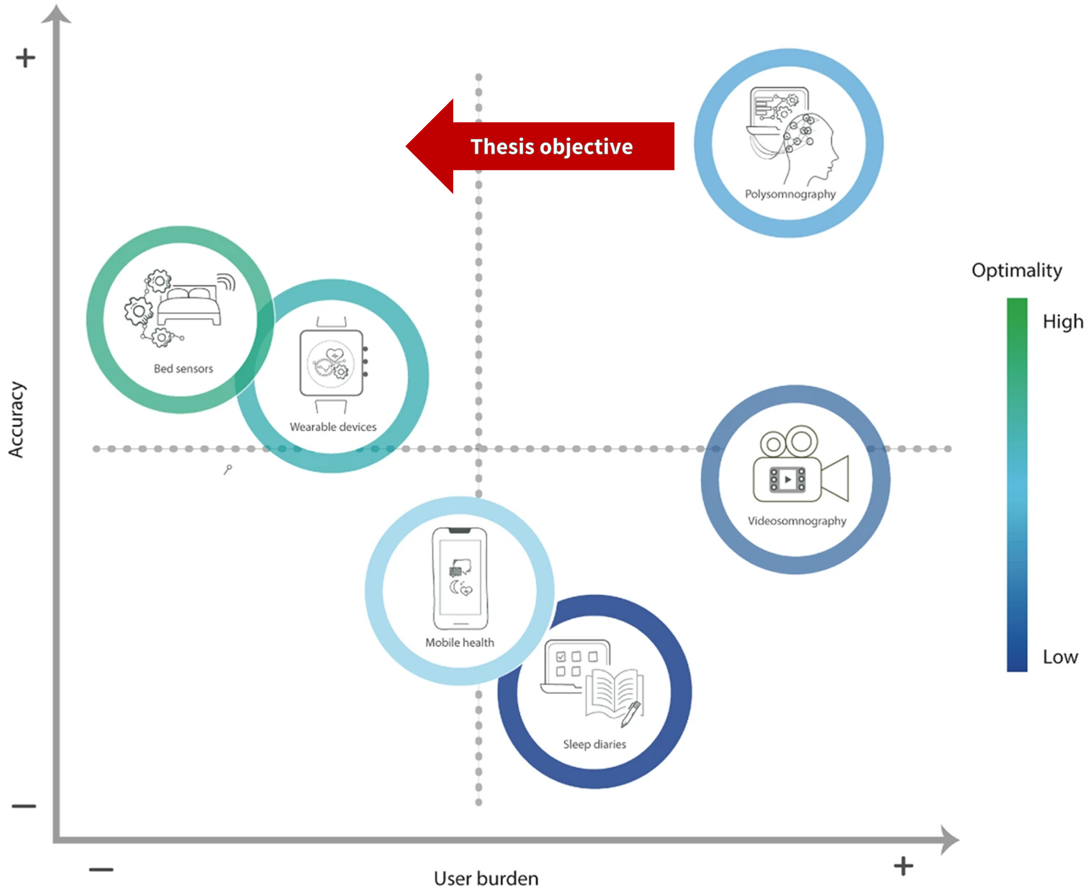


Figure 1.1: Accuracy-usability trade-off for selected methods in sleep analysis. Polysomnography is considered the gold standard in sleep medicine providing high levels of accuracy, while also being very cumbersome and time-consuming. Videosomnography provides lower levels of accuracy, while being equally cumbersome and time-consuming, although this is required for diagnosis of some sleep disorders. User-centered technologies such as health apps and sleep diaries are easier to use than the gold standard methods, but are generally not accurate in capturing sleep metrics. Wearable devices and ubiquitous technologies, such as bed and radiowave sensors, generally have the lowest impact on the user, while providing medium levels of accuracy. This thesis will focus on improving the the gold standard indicated on the figure by moving polysomnography along the red arrow to the left along the user burden axis. In this case, user burden encompasses both the burden to the patient, as well as the burden to the clinician. Adapted from [30] under a Creative Commons Attributes 4.0 International License: `http://creativecommons.org/licenses/by/4.0/`.

Currently, the gold standard of diagnosing sleep disorders is based on manual analysis of sleep patterns following the guidelines published by the American Academy of Sleep Medicine (AASM) and the clinical guidelines in the ICSD [1]. Sleep is recorded in sleep clinics using PSG, which comprises several recording modalities across the body. Patients in the sleep clinic wear and sleep with a heavy and extensive recording setup, which may have an impact on regular sleep patterns [19], [20].

Specific details concerning PSG setup and recording will be described in Section 2.2.

Apart from being time consuming and cumbersome for both patients and clinical personnel, there is growing evidence that manual analysis of sleep patterns suffer from subjectiveness resulting in high inter-scorer variabilities [21]–[29]. In the last decades, there has been an increasing effort to come up with new solutions that can ease and standardize the way sleep data is acquired and analyzed.

These findings on variability will be presented in Section 2.3.

Numerous commercial, industrial and academic interests are focusing on easier methods for acquiring sleep data in the form of mobile health applications and low-cost wearable/nearable devices such as headbands, in-ear EEG, or activity trackers [30]. These devices are interesting from several standpoints, but mostly due to the low user burden compared to conventional PSG. However, wearable devices are not yet applied in clinical practice, due to the limited validation against gold standard methods [31]. This trade-off between usability and user burden versus accuracy is depicted graphically in Figure 1.1 for several methods available for sleep recording and/or quantification.

Nearables *include non-contact devices such as radar-based sensing, and embedded sensors such as bed sensors.*

Similarly, numerous efforts have already been made on the data analysis side to automate the sleep analysis process. Especially the task of automatic sleep stage classification has been the subject of many research papers [32]. With the rising presence of artificial intelligence in medicine, and deep learning in particular [33]–[35], more and more research groups are focusing on applying advanced signal processing and analysis techniques to sleep data. Due to the vast number of different approaches regarding the number of classified sleep stages, feature extraction, classification algorithms, applied datasets, and validation approaches, direct comparison between published findings is a complex task [32], [36]–[39]. However, high quality large-scale studies on automatic methods for sleep analysis was until very recently not dominant in the literature.

1.1 PROBLEM STATEMENT AND RESEARCH HYPOTHESIS

Analysis of sleep is based on manual scoring of PSGs recorded overnight at either a sleep clinic or at home, which are prone to subjective interpretation of scoring rules. Correct identification and analysis of sleep patterns precedes correct diagnosis and thus subsequent treatment of sleep disorders. **The objective of this thesis is**

> *to develop a system based on artificial intelligence, that can assist clinicians in the analysis of sleep studies.*

The aim is to ease the way PSGs are analyzed in the clinic today, but without lowering accuracy. This is depicted graphically in Figure 1.1 by moving the PSG along the red arrow.

Taking into account the motivation, the problem statement is formalized into the following **thesis hypothesis**:

> Advanced biomedical signal processing and machine learning algorithms can be used for efficient, high-performing analysis of sleep studies with regards to

RH 1 sleep stages;

RH 2 sleep events; and,

RH 3 sleep disorders.

1.2 THESIS OUTLINE AND SCIENTIFIC CONTRIBUTIONS

The scientific content of this research is grouped into three research themes each with its own chapter.

Chapter 1 contains the preliminary introduction and motivation to this thesis, and outlines the content and scientific contributions.

Chapter 2 provides necessary clinical background for readers with little previous knowledge in somnology and sleep medicine.

Chapter 3 presents three studies on *automatic sleep stage classification*, the first research theme.

Chapter 4 presents three studies on a model developed for *sleep micro-event detection*, which is the second research theme.

Chapter 5 presents the development of a narcolepsy classification algorithm, based on outputs from one of the sleep stage classification models presented in Chapter 3. This concerns the third and final research theme.

Chapter 6 integrates the findings of this dissertation in a discussion relative to the stated research hypotheses and objectives.

Chapter 7 concludes the thesis by summing up the main research findings.

Chapter 8 outlines some of the future directions and research perspectives in the field of *computational sleep science*.

The following first-author publications have been published, accepted or submitted during my PhD studies and form the scientific basis of this thesis.[1] Preprints and/or published versions of these are supplied in the appendix.

JOURNAL PAPERS

- J. B. Stephansen*, **A. N. Olesen***, M. Olsen, A. Ambati, E. B. Leary, H. E. Moore, O. Carrillo, L. Lin, F. Han, H. Yan, Y. L. Sun, Y. Dauvilliers, S. Scholz, L. Barateau, B. Hogl, A. Stefani, S. C. Hong, T. W. Kim, F. Pizza, G. Plazzi, S. Vandi, E. Antelmi, D. Perrin, S. T. Kuna, P. K. Schweitzer, C. Kushida, P. E. Peppard, H. B. D. Sorensen, P. Jennum, and E. Mignot, "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy", *Nat. Commun.*, vol. 9, p. 5229, 2018. DOI: 10.1038/s41467-018-07229-3
- **A. N. Olesen**, P. Jennum, E. Mignot, and H. B. D. Sorensen, *Automatic sleep stage classification with deep residual networks in a mixed-cohort setting*, 2020, (*under review*)
- **A. N. Olesen**, P. Jennum, E. Mignot, and H. B. D. Sorensen, *A multi-modal sleep event detection algorithm for clinical sleep analysis*, 2020, (*in preparation*)

CONFERENCE PAPERS

- **A. N. Olesen**, P. Jennum, P. Peppard, E. Mignot, and H. B. D. Sorensen, "Deep residual networks for automatic sleep stage classification of raw polysomnographic waveforms", *2018 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Honolulu, HI, USA: IEEE, 2018. DOI: 10.1109/EMBC.2018.8513080
- **A. N. Olesen**, S. Chambon, V. Thorey, P. Jennum, E. Mignot, and H. B. D. Sorensen, "Towards a Flexible Deep Learning Method for Automatic Detection of Clinically Relevant Multi-Modal Events in the Polysomnogram", *2019 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Berlin, Germany: IEEE, 2019, pp. 556–561. DOI: 10.1109/EMBC.2019.8856570
- **A. N. Olesen**, P. Jennum, E. Mignot, and H. B. D. Sorensen, *Deep transfer learning for improving single-EEG arousal detection*, 2020. arXiv: 2004.05111 [cs.CV], (*accepted*, IEEE EMBC 2020)

Furthermore, I have (co-)authored the following publications during my PhD.

JOURNAL PAPERS

- **A. N. Olesen***, M. Cesari*, J. A. E. Christensen, H. B. D. Sorensen, E. Mignot, and P. Jennum, "A comparative study of methods for automatic detection of rapid eye movement abnormal muscular activity in narcolepsy", *Sleep Med.*, vol. 44, pp. 97–105, 2018. DOI: 10.1016/j.sleep.2017.11.1141

[1] * indicates shared first authorship.

- M. Cesari, J. A. E. Christensen, L. Kempfner, **A. N. Olesen**, G. Mayer, K. Kesper, W. H. Oertel, F. Sixel-Döring, C. Trenkwalder, H. B. D. Sorensen, and P. Jennum, "Comparison of computerized methods for rapid eye movement sleep without atonia detection", *Sleep*, vol. 41, no. 10, pp. 1–11, 2018. DOI: 10.1093/sleep/zsy133
- A. Brink-Kjaer, **A. N. Olesen**, P. E. Peppard, K. L. Stone, P. Jennum, E. Mignot, and H. B. Sorensen, "Automatic detection of cortical arousals in sleep and their contribution to daytime sleepiness", *Clin. Neurophysiol.*, vol. 131, no. 6, pp. 1187–1203, 2020. DOI: 10.1016/j.clinph.2020.02.027
- L. Carvelli, **A. N. Olesen**, A. Brink-Kjær, E. B. Leary, P. E. Peppard, E. Mignot, H. B. Sørensen, and P. Jennum, "Design of a deep learning model for automatic scoring of periodic and non-periodic leg movements during sleep validated against multiple human experts", *Sleep Medicine*, vol. 69, pp. 109–119, 2020. DOI: 10.1016/j.sleep.2019.12.032
- A. Ambati, Y.-E. Ju, L. Lin, **A. N. Olesen**, H. Koch, J. J. Hedou, E. B. Leary, V. P. Sempere, E. Mignot, and S. Taheri, "Proteomic biomarkers of sleep apnea", *Sleep*, 2020, (*in press*)

CONFERENCE PAPERS

- A. B. Klok*, J. Edin*, M. Cesari, **A. N. Olesen**, P. Jennum, and H. B. Sorensen, "A New Fully Automated Random-Forest Algorithm for Sleep Staging", *2018 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Honolulu, HI, USA: IEEE, 2018, pp. 4920–4923. DOI: 10.1109/EMBC.2018.8513413

ABSTRACTS

- A. Brink-Kjær, **A. N. Olesen**, C. A. Jespersen, P. E. Peppard, P. J. Jennum, H. B. Sørensen, and E. Mignot, "0142 Automatic Detection of Cortical Arousals in Sleep using Bi-direction LSTM Networks", *Sleep*, vol. 41, no. suppl_1, A55–A56, 2018. DOI: 10.1093/sleep/zsy061.141
- L. Carvelli, **A. N. Olesen**, E. B. Leary, H. Moore, L. D. Schneider, P. E. Peppard, P. J. Jennum, H. B. Sørensen, and E. Mignot, "0323 Design of a Deep Learning Based Algorithm for Automatic Detection of Leg Movements During Sleep", *Sleep*, vol. 41, no. suppl_1, A124–A124, 2018. DOI: 10.1093/sleep/zsy061.322
- K. P. Jacobsen, **A. N. Olesen**, L. Trap, P. E. Peppard, H. B. Sorensen, P. J. Jennum, and E. Mignot, "0328 Automatic Detection of Respiratory Events During Sleep Using Bidirectional LSTM Networks", *Sleep*, vol. 41, no. suppl_1, A125–A126, 2018. DOI: 10.1093/sleep/zsy061.327
- **A. N. Olesen**, P. E. Peppard, H. B. Sorensen, P. J. Jennum, and E. Mignot, "0316 End-to-End Deep Learning Model For Automatic Sleep Staging Using Raw PSG Waveforms", *Sleep*, vol. 41, no. suppl_1, A121–A121, 2018. DOI: 10.1093/sleep/zsy061.315
- **A. N. Olesen**, J. Thybo, S. Chambon, V. Thorey, P. J. Jennum, H. B. Sorensen, and E. Mignot, "0318 Towards A Deep Learning-based Joint Detection Model For Nocturnal Polysomnogram Events", *Sleep*, vol. 42, no. Supplement_1, A130–A130, 2019. DOI: 10.1093/sleep/zsz067.317
- J. Thybo, **A. N. Olesen**, M. Olsen, E. Leary, P. Jennum, H. B. D. Sorensen, and E. Mignot, "Fully Automatic Detection of Sleep-disordered Breathing Events", *Sleep*, 2020, (*in press*)

Finally, I have also written a popular science article about my research titled *Intelligente algoritmer på søvnklinikken* (Intelligent algorithms in the sleep clinic) for the Danish industry magazine *Medicoteknik*.

2 CLINICAL BACKGROUND

What, so everyone's supposed to sleep every single night now? You realize that nighttime makes up half of all time?

— Rick Sanchez
Rick and Morty, season 1, pilot episode

This chapter aims to provide the reader with a basic and preliminary understanding of sleep science. First, the fundamental aspects of sleep as a physiological phenomenon are reviewed. Unless otherwise stated, the context will be concerning sleep in primarily healthy adults. This will be followed by a description of how sleep is recorded, quantified and analyzed in clinical practice. The chapter will conclude with a section on some of the major challenges and difficulties that arise in clinical sleep practice, such as inter- and intra-rater variability, and how this can affect clinical outcomes.

2.1 FUNDAMENTAL ASPECTS OF SLEEP

Sleep is ubiquitous to human life. Our bodies might seem static, but sleep is actually a complex, physiological state comprised of multiple, dynamic processes, that are observable across multiple recording modalities. But although we spend almost a third of our lifetime sleeping, there are still many aspects that are unknown to science.

The general understanding of how our sleep is structured includes two concepts important for this thesis: sleep architecture and sleep events.

SLEEP ARCHITECTURE refers to the structure of sleep, how it is divided into different states based on physiological characteristics, and the dynamics of those states across the night. This can also be called *macro-sleep*, as it concerns the overall macro-structure of our sleep patterns.

SLEEP EVENTS are discrete observations with various characteristics that are distinct for the specific event type. Many such events can happen during sleep, and the duration and scope of these events can vary from short and localized (leg movements, sleep spindles), to long and broad (arousals, apneas). The description and characterization of these events

Table 2.1: Clinical EEG frequency bands.

Rhythm	Brown *et al.* [40]	AASM2020 [43]
Delta	1 Hz to 4 Hz	0 Hz to 3.99 Hz
Theta	4 Hz to 8 Hz	4 Hz to 7.99 Hz
Alpha	8 Hz to 14 Hz	8 Hz to 13 Hz
Beta	15 Hz to 30 Hz	$>$ 13 Hz
Gamma	30 Hz to 120 Hz	n.d.
SWA	0.5 Hz to 4 Hz	0.5 Hz to 2.0 Hz
LAMF	n.d.	4 Hz to 7 Hz

n.d., not defined; EEG: electroencephalography; SWA: slow wave activity; LAMF: low amplitude, mixed frequency.

can also be called *micro-sleep*, but this term is also sometimes applied to sleep architecture on a small time-scale. In this thesis, I will refer to this concept as either *micro-sleep events* or just *sleep events* for short.

2.1.1 *Sleep architecture*

On average, normal sleep in adult humans lasts between 7-9 hours per night with substantial variability between persons. During this period, the brain and body cycle between alternating *sleep stages*, which can be categorized into a state of drowsiness or semi-conscious wakefulness (W), a rapid eye movement (REM) sleep stage, and three non-rapid eye movement (NREM) stages, non-rapid eye movement stage 1 (N1), non-rapid eye movement stage 2 (N2), and non-rapid eye movement stage 3 (N3). The main distinction between sleep stages comes from the amplitude and spectral content of the brain signals as measured using EEG. For example, wakefulness and REM sleep are associated with high frequency, low amplitude content such as theta or alpha rhythm activity, while the NREM sleep stages are associated with low frequency, high amplitude content in the delta rhythm range.

However, certain brain stages are also characterized by the presence of certain micro-structure events with very distinct morphologies, such as sleep spindles or K-complexes in the EEG, or REMs which are recorded with EOG [40]–[42]. Muscle activity, which is typically recorded using EMG of the submentalis and anterior tibialis muscles, can also be used to distinguish between sleep stages.

The submentalis is located below the chin, while the anterior tibialis is located on the shin.

The following sections review the major electrophysiological characteristics for the five sleep stages currently defined by the AASM, which is summarized in Table 2.2. A graphical overview of typical signal content in the EEG, EOG and EMG for the five sleep stages is shown in Figure 2.1.

2.1.1.1 *Wakefulness*

Spanning from a full awareness state to a quiet awakening or drowsiness state, this stage generally accounts for about 5 % of the total time in bed from lights out to lights on in healthy adults. In this stage, the brain typically exhibits low amplitude, high frequency content in small areas and more wide-spread theta rhythms. During quiet awakening, these theta rhythms increase in the frontal area of the brain, while alpha rhythms are dominant

EEG rhythms in the alpha range or higher.

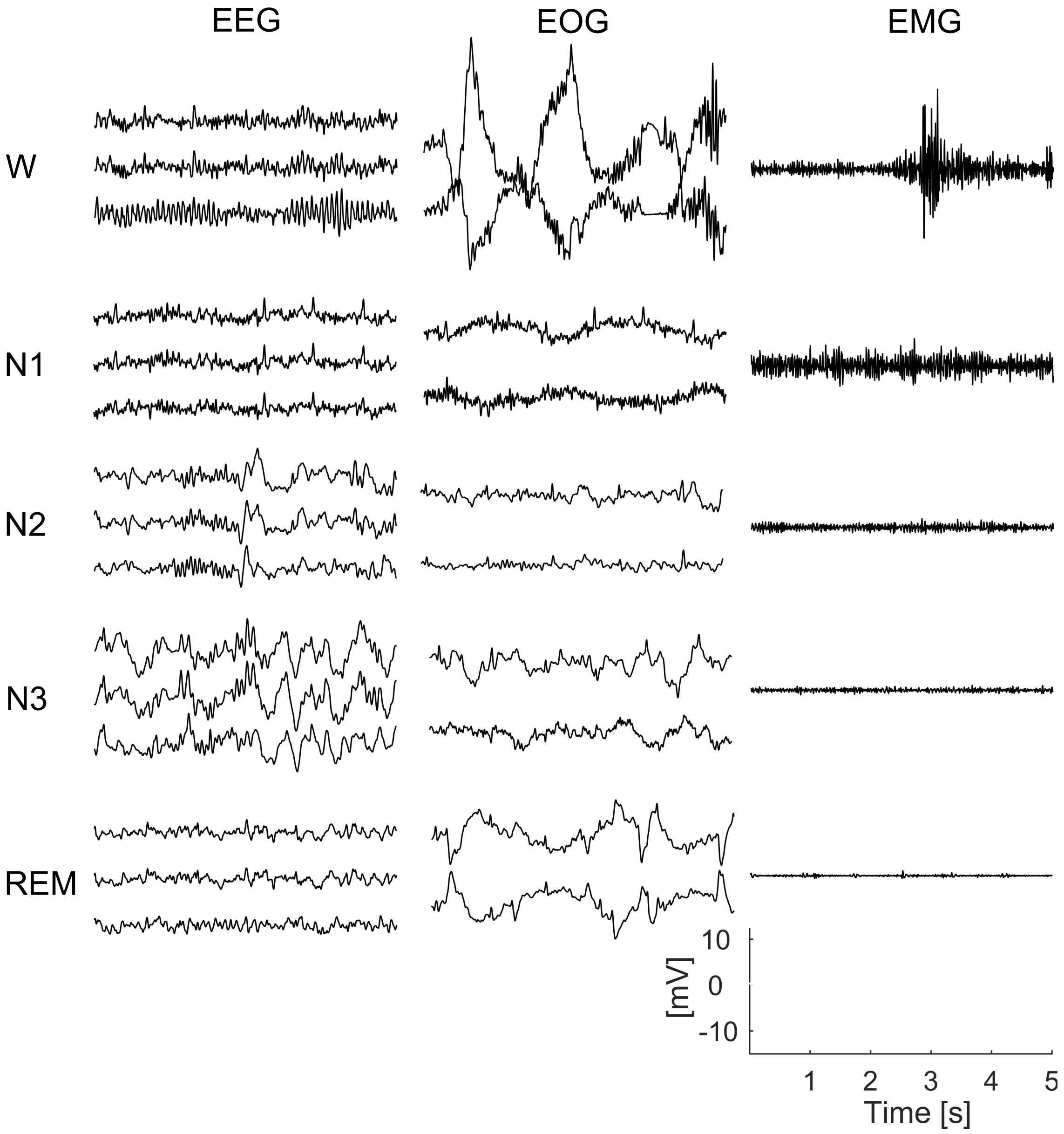


Figure 2.1: Typical waveforms encountered in different sleep stages. The left column displays the progression from low amplitude, mixed frequency content such as alpha rhythms in W, towards low frequency, high amplitude delta rhythms in the deep N3 sleep. In the middle is shown REMs and reading eye movements in W, and SEMs or no activity in NREM stages. Due to their proximity to the frontal area, delta rhytms might be visible in the EOG during N3. The right-most column shows the EMG amplitude progression from high in W to low in N3. The bottom row shows the paradoxical nature of REM sleep with LAMF content and theta rhythms in the EEG, REMs in the EOG, and atonia in the EMG. Reprinted with permission from [44].

Table 2.2: Sleep stage characteristics according to AASM2020 [43].

Sleep stage	Description
W	EEG: alpha (eyes closes), theta (eyes open) EOG: REMs, reading eye movements EMG: high, but variable, amplitude
N1	EEG: LAMF content, theta rhythms EOG: SEMs EMG: variable amplitude, usually lower than in W
N2	EEG: K-complexes, sleep spindles EOG: No activity, or SEMs EMG: variable amplitude, usually lower than in N1
N3	EEG: SWA, large amplitude delta rhythms EOG: No activity, or SEMs EMG: variable amplitude, usually lower than in N2
REM	EEG: LAMF content, theta rhythms, sawtooth waves, PGO waves EOG: REMs EMG: muscle atonia, phasic bursts

LAMF: low amplitude, mixed frequency; PGO: ponto-geniculo-occipital; REM: rapid eye movement; SEM: slow eye movement.

over the occipital region, especially when the eyes are closed. With eyes open, this stage is characterized by eye blinking, reading eye movements and REMs. The chin muscle tone is typically high with unspecific amplitude.

2.1.1.2 *NREM sleep*

Up until 2007, NREM sleep was divided in four stages: S1, S2, S3, and S4 (slow wave sleep). The latter two was merged in AASM2007 into N3.

Sleep in humans generally commences when a person progresses from W to one of the three stages of NREM. They generally constitute approximately 2 % to 5 %, 45 % to 55 %, and 13 % to 23 % of the total sleep time (TST) for the first, second, and third NREM stage, respectively. The order of N1, N2, and N3 represents a continuum of the depth of sleep, which is primarily constituted by a progressive slowing of the EEG activity from low amplitude, high frequency alpha and theta rhythms, to low frequency delta rhythms with large amplitude. Furthermore, another major indicator of deepening sleep is the increasing arousal thresholds associated with the progression from N1 to N3 [40], [42].

N1 This is the first stage encountered during normal sleep, and is generally considered to be a transitional stage between drowsiness, or lighter sleep, and deeper sleep. It is characterized by mixed frequency content, as the alpha rhythms are progressively reduced and replaced with theta rhythms. Additionally, small sleep events called *vertex sharp waves* also appear in the EEG during this stage. REMs and reading eye movements are replaced with slow eye movements and the muscle tone is reduced compared to W.

Vertex sharp waves are sharply contoured waveforms with a very short duration of less than 0.5 s

N2 Although theta rhythms from N1 continue, brain activity in N2 exhibits discrete sleep events like *sleep spindles* and *K-complexes*. Eye movement activity

These events will be described in Section 2.1.3 is not typically seen in this stage, while muscle activity is further reduced from that of N1.

N3 The last of the three NREM, the N3 stage is also known as *deep sleep*. In this stage, the EEG content shifts towards more pronounced delta wave activity, especially slow wave activity (SWA) larger than 75 µV in the frontal regions. Eye movement activity is typically not seen, but SWA intrusion artifacts from the frontal EEG can sometimes be seen. The muscle activity is low, and the muscle tone typically has an amplitude between that of N2 and REM.

2.1.1.3 *REM sleep*

The defining characteristic of this stage of sleep is the appearance of REM activity observed in the EOG, desynchronized low amplitude, mixed frequency (LAMF) EEG activity, and suppression of skeletal muscle activity (atonia) due to brainstem-mediated inhibition of the alpha motor-neurons [42], [45]. This stage typically accounts for 20 % to 25 % of the TST [42], and is associated with the majority of dreaming activity, although evidence suggests dreaming also occurs in NREM [40], [45]–[48], albeit at a less vivid level [49].

Conjugate, irregular, sharply peaked eye movements with an initial deflection lasting less than 500 ms [43]

Some studies indicate that REM sleep is comprised of two distinct microstates [45]: a *phasic* state characterized by the appearance of ponto-geniculo-occipital (PGO) and *sawtooth* waves, REMs and muscle twitches [50]; and a *tonic* state in between phasic states that is characterized by LAMF and theta rhythms, and a higher arousal threshold [51]. Studies on the effects of auditory stimulation and event related potentials in REM microstates confirm the heterogeneity of REM sleep [52]–[54]. This is underlined by clinical studies on REM sleep behaviour disorder (RBD) patients, which suggest a distinction in the level of activation of the sensory motor system during tonic and phasic REM [55]–[57].

Characteristic 2-4 Hz rhythmic oscillations during REM sleep that have a triangular shape and maximal amplitude at frontal and central derivations.

2.1.2 *Neurobiological control of sleep*

2.1.2.1 *The wake-sleep switch*

The *wake-sleep switch* consists of several neuronal populations in the areas of the upper brainstem, hypothalamic, and forebrain responsible for promoting wakefulness and sleep, as shown in Figure 2.3 [41].

Two major pathways promoting wakefulness are located in the upper brainstem. Figure 2.3A shows cholinergic neurons (cyan) pedunculopontine tegmental nucleus (PPT) projecting to the thalamus, lateral hypothalamic area (LHA), and basal forebrain (BF) from the laterodorsal tegmental nucleus (LDT), thereby driving extensive cortical activation. The monoaminergic and glutaminergic neurons (green) in the locus coeruleus (LC), parabrachial nucleus (PB), precoeruleus nucleus (PC), dorsal raphe nucleus (DRN), ventral periaqueductal gray (vPAG), and tuberomammillary nucleus (TMN) project to the forebrain, cerebral cortex, thalamic and hypothalamic areas. A special population of neurons in the LHA promotes wakefulness by directly exciting both wakefulness pathways as well as the BF and cerebral cortex by secreting hypocretin (hcrt) [59].

Two research teams simultaneously discovered these neuropeptides and gave them the names hypocretin and orexin, respectively.

The main pathway for sleep promotion, shown in purple in Figure 2.3B, originates in neurons in the ventrolateral preoptic nucleus (VLPO) and median preoptic nucleus (MnPO) areas of the hypothalamus and inhibits the pathways promoting wakefulness. In turn, these pathways can likewise

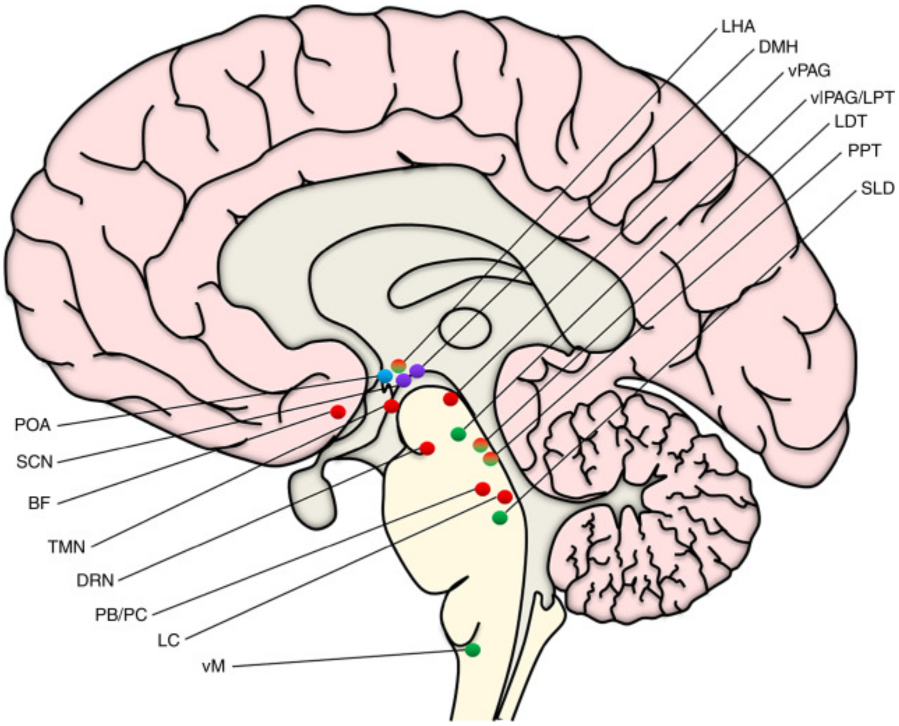


Figure 2.2: Locations of the brain centers involved in sleep-wake control. The predominant role of each center is color-coded: red for arousal, blue for sleep, green for REM sleep, purple for circadian regulation, and multicolored for mixed activity. BF: basal forebrain; DMH: dorsomedial hypothalamic nucleus; DRN: dorsal raphe nucleus; LC: locus coeruleus; LDT: laterodorsal tegmental nucleus; LHA: lateral hypothalamic area; LPT: lateral pontine tegmentum; PB: parabrachial nucleus; PC: precoeruleus nucleus; POA: preoptic area; PPT: pedunculopontine tegmental nucleus; SCN: suprachiasmatic nucleus; SLD: sublaterodorsal nucleus TMN: tuberomammillary nucleus; vM: ventral medulla; vlPAG: ventrolateral periaqueductal gray; vPAG: ventral periaqueductal gray. Reprinted from [58] with permission from Elsevier.

inhibit the sleep-promoting activity in the VLPO and MnPO areas, as shown in Figure 2.3C.

This mutual inhibitory activity is the basis for the wake-sleep switch acting like an electronic switch. When one group of neurons gains a slight advantage over the other, a rapid decrease in neuronal firing activity on the losing side ensures a fast and complete transition between the two states [41], [58].

2.1.2.2 *The REM-NREM switch*

Similar to the wake-sleep switch, the REM-NREM switch consists of several groups of neurons in the brainstem and hypothalamic areas, as shown in Figure 2.4 [41].

Two populations of importance, shown in Figure 2.4A, include a group of neurons in the sublaterodorsal nucleus (SLD) and PC (red) that fires most actively during REM, and a group in the ventrolateral periaqueductal gray (vlPAG) and lateral pontine tegmentum (LPT) (gold) that promotes NREM sleep by inhibiting REM-on neurons.

These two populations are acted upon by several other neurotransmitter systems as shown in Figure 2.4B. One of these consists of noradrenergic neurons in the LC and serotonergic neurons in the DRN (green), which inhibits REM sleep by exciting REM-off neurons and inhibiting REM-on neurons. Conversely, a system of cholinergic neurons (aqua) in the LDT

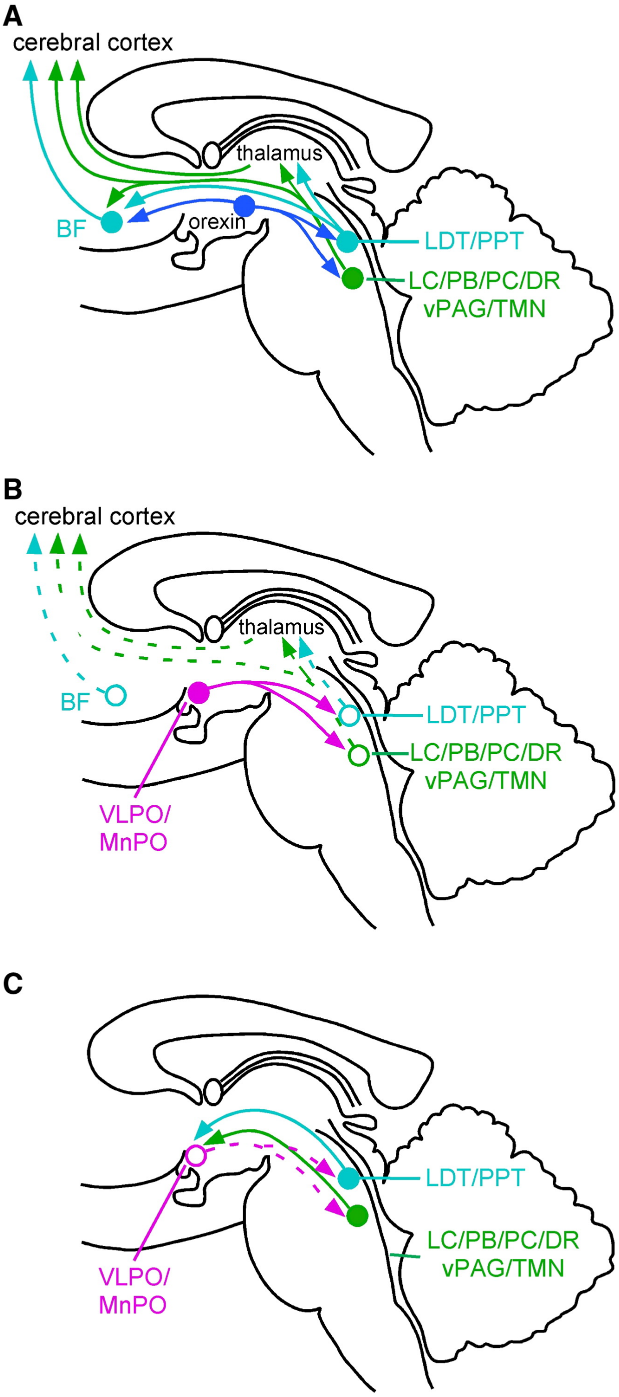


Figure 2.3: Structures involved in the wake-sleep switch. **A** The two main wake-promoting pathways originating in the upper brainstem. Cholinergic neurons (aqua) in the LDT and PPT project to the thalamus and BF, while predominantly monoaminergic neurons (green) in the LC, PB, PC, DR, vPAG, and TMN project directly to the hypothalamus, BF and cerebral cortex. Hypocretin-producing neurons in the hypothalamus reinforce activation in both pathways, while directly innervating the BF. **B** The main sleep-promoting pathways originate in the VLPO and MnPO areas (purple) of the hypothalamus and inhibit the activity of the wake-promoting networks in the upper brainstem. **C** The wake-promoting networks in the brainstem in turn can also inhibit the activity of the sleep-promoting networks in the VLPO/MnPO area. BF: basal forebrain; DR: dorsal raphe; LC: locus coeruleus; LDT: laterodorsal tegmental nucleus; MnPO: median preoptic nucleus; PB: parabrachial nucleus; PC: precoeruleus nucleus; PPT: pedunculopontine tegmental nucleus; TMN: tuberomammillary nucleus; VLPO: ventrolateral preoptic nucleus; vPAG: ventral periaqueductal gray. Reprinted from [41] with permission from Elsevier.

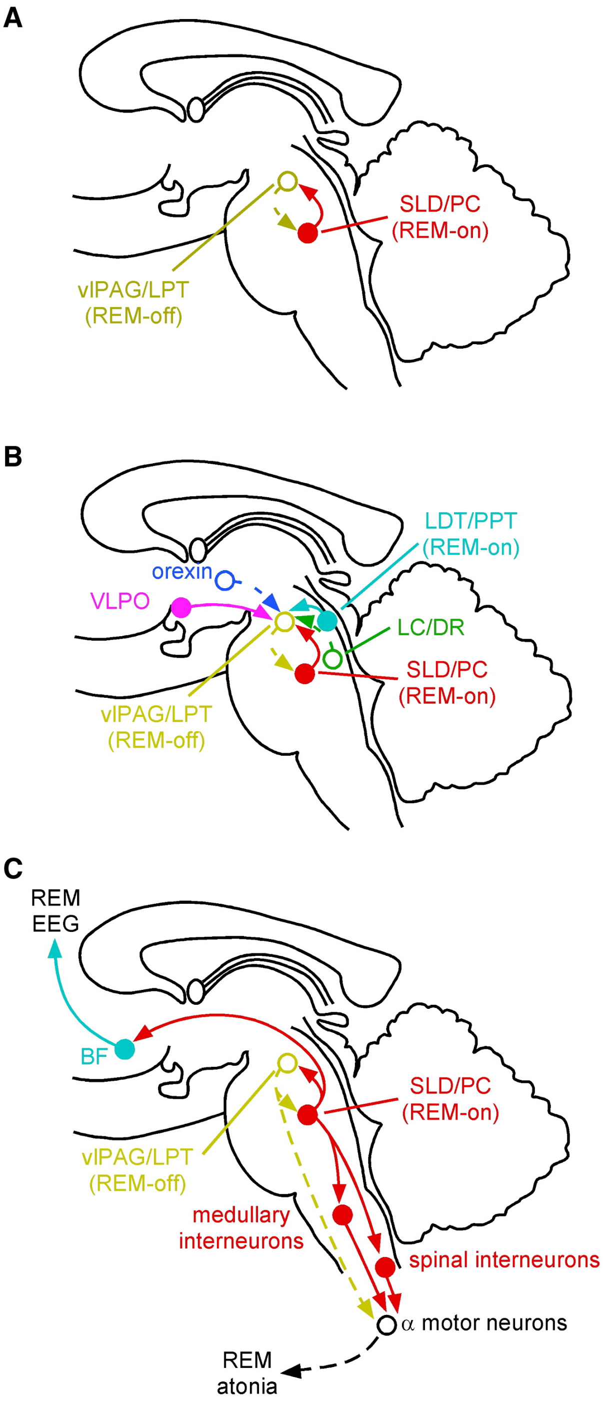


Figure 2.4: Structures involved in the REM-NREM switch. **A** Two mutually inhibitory GABAergic neuron groups in the brainstem form a switch for controlling transitions between REM and NREM sleep. Projections from the SLD and PC (red) promote REM, while REM-off projections from the vlPAG and LPT (gold) are most active during NREM and thus inhibit REM-producing activity in the SLD/PC nuclei. **B** The REM-NREM switch is also modulated by other neurotransmitter systems. The noradrenergic (green) neurons in the LC and serotonergic DR inhibit REM sleep by exciting the REM-off neurons and inhibiting REM-on neurons. Cholinergic (aqua) neurons in the LDT and PPT nuclei promote REM sleep by the opposing actions on the same neuron groups. VLPO neurons promote REM sleep by inhibiting the REM-off neurons, while orexin-producing neurons produce the opposite effect. **C** During REM sleep, glutaminergic neurons in the SLD promote REM atonia in the skeletal muscles by way of inhibitory interneurons in the spinal cord and medulla, which act on the α motor neurons. Furthermore, glutamergic neurons in the PB/PC promote desynchronized EEG via BF neurons. BF: basal forebrain; DR: dorsal raphe; LC: locus coeruleus; LDT: laterodorsal tegmental nucleus; LPT: lateral pontine tegmentum; PC: precoeruleus nucleus; PPT: pedunculopontine tegmental nucleus; SLD: sublaterodorsal nucleus; vlPAG: ventrolateral periaqueductal gray; VLPO: ventrolateral preoptic nucleus. Reprinted from [41] with permission from Elsevier.

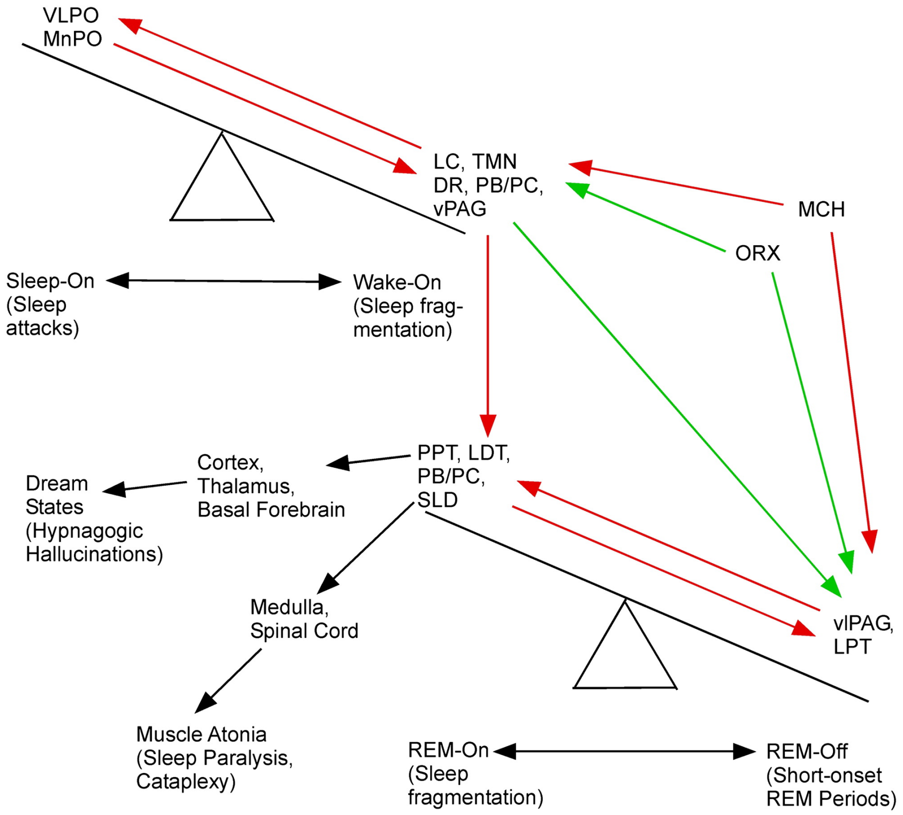


Figure 2.5: Schematic of the flip-flop model. Clinical manifestations of wrongful firing of neurons are shown in parentheses. DR: dorsal raphe; MnPO: median preoptic nucleus; LC: locus coeruleus; LDT: laterodorsal tegmental nucleus; LPT: lateral pontine tegmentum; MCH: melanin-concentrating hormone; ORX: orexin (hypocretin); PB: parabrachial nucleus; PC: precoeruleus nucleus; PPT: pedunculopontine tegmental nucleus; SLD: sublaterodorsal nucleus; TMN: tuberomammillary nucleus; vlPAG: ventrolateral periaqueductal gray; VLPO: ventrolateral preoptic nucleus; vPAG: ventral periaqueductal gray. Reprinted from [41] with permission from Elsevier.

and PPT promotes REM sleep by the opposite mechanisms. Neurons in the VLPO (purple) also promote REM sleep by inhibiting REM-off neurons, while hcrt-(orexin-)producing neurons in the LHA inhibit REM sleep by exciting REM-off neurons.

Lastly, the glutaminergic neurons in the SLD promote REM atonia by exciting inhibitory interneurons in the medulla and spinal cord, as shown in Figure 2.4C.

2.1.2.3 *The flip-flop model of sleep*

It can be appreciated from the previous sections, that sleep and wake states are under the regulation of extremely complex mechanisms in the brainstem, thalamus, hypothalamus, and forebrain. These mechanisms can be grouped into the wake-sleep and REM-NREM promoting networks, which, in turn, can be combined into the *flip-flop* model of sleep regulation [41], [60], [61].

A schematic shown in Figure 2.5 illustrates both the complex neuronal interplay, as well as how two neurotransmitters play crucial roles in sleep homeostasis. One is the neuropeptide hypocretin, which acts to inhibit REM entry by exciting the wake-promoting pathways and the REM-off nuclei, while melanin-concentrating hormone (MCH) has the opposite effect of hypocretin.

Also shown in Figure 2.5 are clinical manifestations (in parentheses) associated with destabilized switching due to a lack of hypocretin.

Lack of hypocretin-producing neurons is the hallmark sign of NT1.

2.1.3 *Micro-events durings sleep*

2.1.3.1 *Arousals*

Arousals are defined as abrupt shifts in EEG frequencies towards alpha, theta, and beta rhythms for at least 3 s with a preceding period of stable sleep of at least 10 s. During REM sleep, where the background EEG shows similar rhythms, arousal scoring requires a concurrent increase in chin EMG lasting at least 1 s. [51]

2.1.3.2 *Movements of the extremities*

LMs should be scored in the leg EMG channels, when there is an increase in amplitude of at least 8 µV above baseline level with a duration between 0.5 s to 10 s. A periodic leg movement (PLM) series is then defined as a sequence of 4 LMs, where the time between LM onsets is between 5 min to 90 min.

2.1.3.3 *Respiratory disturbances*

Apneas are generally scored when there is a complete (⩾90 % of pre-event baseline) cessation of breathing activity either due to a physical obstruction (obstructive apnea) or due to an underlying disruption in the central nervous system control (central apnea) for at least 10 s. When the breathing is only partially reduced (⩾30 % of pre-event baseline) and the duration of the excursion is ⩾10 s, the event is scored as a hypopnea if there is either a ⩾4 % oxygen desaturation or a ⩾3 % oxygen desaturation coupled with an Ar.

2.2 RECORDING AND QUANTIFYING SLEEP

2.2.1 *Polysomnography*

The principal tool available to sleep physicians and technicians for analysis of sleep patterns is the *polysomnography* (PSG). This is often the first study performed on patients referred to a sleep clinic, and consists of the continuous and concurrent recording of several physiological variables as electrophysiological signals. The primary signals of interest are brain activity (EEG), eye movements (EOG), chin and leg muscle activity (EMG), heart activity (electrocardiography (ECG)), respiratory effort (thoracoabdominal inductance plethysmography belts, RIP), nasal pressure, oral airflow, and blood oxygen saturation (pulse oximetry). Sleep experts manually analyze the contents of these signals in order to score sleep stages and annotate sleep events based on a standardized set of guidelines published by the AASM [43]. These guidelines also contain technical recommendations for recording sleep studies, such as electrode placements, minimal sampling frequencies and specific filter settings. Table 2.3 lists an overview of technical specifications for commonly recorded signals as recommended by the AASM.

A common procedure for analysis of sleep studies involves multiple passes through each PSG study. For example, a first pass could be to score every consecutive segment of 30 s data as one of the five sleep stages. A second

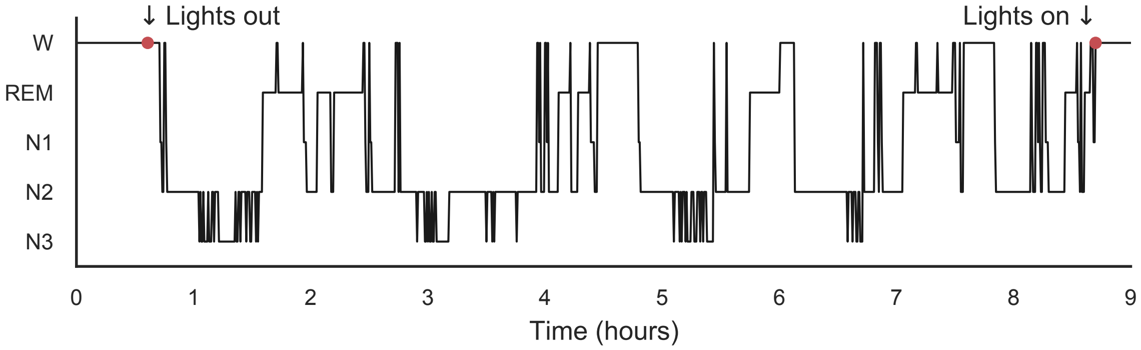


Figure 2.6: Example of a hypnogram from a PSG study of an 82 year old male subject.

Table 2.3: Technical specifications for recording commong signals in PSGs according to AASM2020 standards [43].

Signal	Recommended recording setup	Min. f_s	Filter
EEG	F4-M1, C4-M1, O2-M1 (required)	200 Hz	0.3–35 Hz
	F3-M2, C3-M2, O1-M2 (backup)		
EOG	E1-M2, E2-M2 (required)	200 Hz	0.3–35 Hz
	E1-M1, E2-M1 (backup)		
EMG	Chin2-ChinZ (required, chin EMG)	200 Hz	10–100 Hz
	Chin1-ChinZ (backup, chin EMG)		
	Bipolar derivation (required, leg EMG)		
ECG	modified Lead II derivation (required)	200 Hz	0.3–70 Hz

AASM describes a specific requirement as well as a backup in case of failure for each signal. Minimal f_s lists the minimally acceptable sampling frequency per signal, but the recommendations are higher in order to better capture waveform morphology. Filter settings describe recommended bandpass filter settings.

pass could be to score respiratory events, arousals and leg movements, etc. The product of these passes is a sleep study report, which summarizes the findings into a hypnogram and associated PSG variables, including total sleep time (TST), sleep latency (SL), REM latency (RL), wake after sleep onset (WASO), and percentage of time spent in the different sleep stages. Key indices describing the amount of sleep events are also calculated for each study, such as the arousal index (number of arousals per hour of sleep, ArI), apnea-hypopnea index (number of apneas and hypopneas per hour of sleep, AHI), and the periodic limb movements in sleep index (number of peridoc limb movement series in sleep per hour of sleep, PLMSI).

2.3 Challenges in scoring sleep studies

Significant human bias can enter into the analysis of sleep studies by virtue of the process being performed manually. Several studies have shown significant *inter*- and *intra*-rater variability primarily in the case of sleep stage scoring, but some studies have also investigated the reliability of scoring arousals and respiratory events for sleep-related breathing disorders. This variability can be caused by several factors:

Interrater variability refers to the variation in scoring that happens between experts on the same study, while intrarater variability refers to the variability in scoring when a single expert scores the same study more than once.

1. *Imprecise scoring guidelines.* Some have argued that extensive training is required to minimize the subjective component in sleep stage scoring, and that the optimal training requires participation in concensus scoring rounds [62].

2. *Presence of disease or other sleep disorders.* Many neurodegenerative diseases exhibit symptoms of disturbed sleep as the neurodegeneration progresses to the centers in the brain stem responsible for control of sleep and wakefulness [63]. Similarly, central hypersomnias can also exhibit fragmented sleep. NT1, for instance, shows increased fragmentation of sleep, because the hypocretin-producing cells in the latero-posterior hypothalamus responsible for stabilizing the wake-sleep and REM-NREM pathways are missing [64]. Since current scoring guidelines are based on clinical experience in healthy subjects exhibiting normal sleep patterns, the scoring of sleep patterns becomes more difficult in this context.

Central disorders of hypersomnolence, hypersomnias, *are a group of sleep disorders characterized by excessive daytime sleepiness not caused by disturbed nocturnal sleep or irregular circadian rhythms [1]*

3. *True errors.* These can occur when annotations are correctly made, but entered wrongly into a computer system or report. However, these types of errors are difficult to measure in practice.

A separate, but equally critical issue in scoring sleep studies concerns recording of the study itself by means of electrical equipment. Typically, the PSG is performed as described in Section 2.2.1 with many electrodes placed on the body to the discomfort of the wearer potentially disrupting regular sleep patterns. For this reason, efforts are being made in industry and research to develop non-intrusive, low-impact recording devices, such as headbands, or in-ear EEG [65], [66].

Depending on the target variable under investigation, reliability and variability can be measured with different metrics. The following sections will describe some of the studies that have investigated inter- and intra-scorer reliability for various sleep analysis objectives.

2.3.1 *Sleep stage scoring*

Norman *et al.* found the average epoch by epoch agreement between five experienced PSG technicians representing different clinics to be 73 % in a dataset containing 62 PSGs [24]. Furthermore, they also found this agreement to vary with phenotype, as the average agreement in a normal subset was higher than for a subset consisting of patients with sleep disordered breathing (76 % in the normal subset vs. 71 % in the SDB subset).

Later studies also found significant variability in expert agreements when comparing different patient groups. Notably, Danker-Hopfe *et al.* investigated interrater reliability between experienced technicians in eight different sleep clinics in Europe in a sample of 196 recordings from 98 patients exhibiting different disorders, such as depression, general anxiety disorder with and without insomnia, Parkinson's disease, sleep apnea and periodic leg movements in sleep disorder [23]. They found that although the overall agreement between experts as measured by Cohen's κ was 0.6816, there was a statistically significant difference between patient groups, where the median κ ranged from 0.6138 in patients with Parkinson's disease to 0.8176 in patients with generalized anxiety disorder. Other studies have found no statistically significant differences in the overall agreement between healthy controls, patients with sleep apnea/hypopnea syndrome, and patients with narcolepsy,

Cohen's κ is a measure of the observed agreement between two agents taking into account chance agreement.

when comparing scorers from Berlin and Beijing [67]. They did, however, find statistically significant differences in the stage-specific agreements between patient groups.

Recent large scale studies on interscorer agreement found that the average consensus-agreement is approximately 83 % with the overall stage-specific agreement ranging from 63 % for N1 to 91 % for REM [26]. Although the authors recognize that their results are heavily biased towards agreements in the N2 stage as this accounts for almost 60 % of the total number of epochs, this percentage is in agreement with clinical experience and reflects the amount of N2 in a typical sleep study.

Although human subjective bias is also a factor, the vast majority of interscorer differences originate from equivocal epochs that have equal probability of being assigned to two stages. Younes, Raneri, and Hanly found that disagreements were most common between W and N1, N2 and N3, and N1 and N2 [28], and indeed several studies have found that scoring N1 and N3 sleep is especially difficult [23], [26], [29], [67].

2.3.2 *Arousals*

The majority of studies on reliability of arousal scoring are based on criteria from the American Sleep Disorders Association (ASDA) [68]. One study comparing several different criteria for arousal scoring found an intraclass correlation coefficient (ICC) of 0.84 using the ASDA criteria [69]. Experts scoring arousals shorter than 3 s with an ICC between 0.19 and 0.37 were found less reliable, while the addition of increased EMG activity as a criteria in addition to the ASDA criteria increased the ICC to 0.92. Another study, however, did not improve the already-high ICC of 0.98 when supplementing the ASDA criteria with increased EMG activity [70].

the intraclass correlation coefficient is a descriptive statistic for characterizing agreement between data that can be naturally organized in groups.

Another factor to be considered in the reliable scoring of arousals is the placement of the arousal in the sleep continuum. Drinnan *et al.* investigated the impact of sleep stage on arousal scoring and found the highest Cohen's κ value for arousals scored in slow wave sleep [21]. This sleep stage exhibits delta and SWA EEG rhythms with high amplitude and low frequency, which is easier to contrast with the shift to high frequency EEG content typically associated with arousals.

Other types of cues visible in the PSG are the presence of autonomous findings such as increased heart rate visible in the ECG, or increased respiratory effort. The latter is evident in a study investigating arousal scoring reliability in 17 obstructive sleep apnea (OSA) patients using ASDA criteria. An event-by-event scoring agreement of 91 %, which dropped significantly to 59 % when removing the respiratory signals was found [71].

While reliability of scoring arousals according to the updated AASM2007 criteria remains severely understudied, Magalang *et al.* reported an intraclass correlation coefficient for the arousal index of 0.68 (95 % CI: [0.49 – 0.85]) in 15 PSGs scored by nine technicians from unique sleep clinics according to AASM2007 criteria [25].

These reported results are based solely on the values of the arousal index per study, and as such, two scorers can potentially score completely different arousals for a specific PSG, while still having good agreement between them, since their scored arousal index values are similar. Furthermore, scoring only index values for a study do not reflect important underlying characteristics of the arousal events. These characteristics could include event morphology and variability in each recorded modality, as well as variations in duration, spectral content, amplitude, *etc.*.

2.3.3 *Sleep disordered breathing*

Whitney *et al.* investigated inter- and intra-scorer reliability in three technicians for 20 randomly selected PSG recordings from the Sleep Heart Health Study (SHHS) cohort using various definitions of respiratory disturbance indices (RDIs) with or without arousals, and oxygen desaturation levels from 2 % to 5 % [22]. The authors found that the technicians were in high agreement when scoring respiratory events with oxygen desaturation levels present indicated by an ICC between 0.90 and 0.99. However, this reliability dropped to moderate agreement when oxygen levels were not part of the scoring (ICC of 0.77), and when neither oxygen levels or arousals were included (ICC of 0.74).

The study by Magalang *et al.* also investigated the agreement in scoring respiratory events. The authors reported an ICC for the apnea-hypopnea index (AHI) of 0.95 (95 % CI: $[0.91-0.98]$) which indicates a very strong agreement between centers.

Rosenberg and Van Hout studied the degree of agreement between more than 3600 scorers when asked to identify whether a 30 s epoch contained either an obstructive, mixed or central apnea; a hypopnea; or no event at all [27]. They found that overall agreement was 93.3%, although this was caused by a very high agreement of 97.4% on epochs with no event present. For epochs with a majority vote of having a respiratory event present, overall agreement was 88.4%. The study also showed that disagreements between apneas and hypopneas, and between different types of apneas were common.

However, as with the arousal scoring, neither the RDI nor the AHI take into account the exact location of respiratory events, which means that two technicians in theory could be in perfect agreement when comparing these values even though they did not score any of the same events.

2.4 CHAPTER SUMMARY

This chapter has introduced some of the fundamental aspects in sleep science.

The concepts of *macro-* and *micro-*sleep were introduced. The former was elaborated upon with descriptions of the five sleep stagescurrently recognized by the AASM, and how they together form a description of the sleep architecture called a hypnogram.

The chapter described the neurobiological mechanisms and complex interplay between cell nuclei in the brainstem responsible for sleep regulation. For example, the wake-sleep and REM-NREM switches consist of several neuron groups in the basal forebrain, hypothalamic areas, and brainstem. Their mutual interactions responsible for the fast and complete transitions between wake and sleep, and REM-on and REM-off periods, can be conceptualized in the flip-flop model, which summarizes the roles of the different cell groups and neurotransmitter systems.

Although, this chapter did not elaborate on various malfunctions of the sleep regulatory systems in the brainstem, later chapters will touch on specific sleep disorders where applicable, such as narcolepsy in Chapter 5.

The final sections in this chapter touched upon some of the issues facing clinical sleep medicine. Specifically, the reproducibility of scoring sleep studies, the intra- and inter-scorer reliability of scoring sleep stages, arousals, and sleep disordered breathing events were presented and discussed. The difficulty in providing accurate and reproducible results with clinical outcomes is specifically a motivating factor for this thesis; namely the development of robust systems to automatically process and analyze clinical sleep studies.

Part II

RESEARCH

3 SLEEP STAGE CLASSIFICATION

`What is my purpose?`
You pass butter.
`Oh my God.`

— Butter Robot to Rick Sanchez
Rick and Morty, season 1, episode 9

This chapter presents methods and main findings in two published research papers and one manuscript currently under review regarding automatic methods for sleep stage classification.

First, the problem of automated sleep stage classification is presented and associated research questions are formulated. Then, an initial version of the multi-modal automatic sleep stage classification (MASSC) model based on end-to-end deep learning is presented along with the results published in [72], which is followed by the findings from applying an updated version of the model in a multi-cohort experimental setting. Afterwards, the Stanford Technology Analytics and Genomics in Sleep (STAGES) model for sleep stage classification originally published in [73] is presented along with results compared to multiple scorers. The chapter will conclude with a summary of the main findings in Section 3.5

Parts of this chapter have been modified from their original publications.

- Section 3.2 has been modified from
  A. N. Olesen, P. Jennum, P. Peppard, E. Mignot, and H. B. D. Sorensen, "Deep residual networks for automatic sleep stage classification of raw polysomnographic waveforms", *2018 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Honolulu, HI, USA: IEEE, 2018. DOI: 10.1109/EMBC.2018.8513080[1].

- Section 3.3 is based on
  A. N. Olesen, P. Jennum, E. Mignot, and H. B. D. Sorensen, *Automatic sleep stage classification with deep residual networks in a mixed-cohort setting*, 2020, (*under review*).

[1] ©2018 IEEE

- Section 3.4 has been modified from
  J. B. Stephansen*, **A. N. Olesen***, M. Olsen, A. Ambati, E. B. Leary, H. E. Moore, O. Carrillo, L. Lin, F. Han, H. Yan, Y. L. Sun, Y. Dauvilliers, S. Scholz, L. Barateau, B. Hogl, A. Stefani, S. C. Hong, T. W. Kim, F. Pizza, G. Plazzi, S. Vandi, E. Antelmi, D. Perrin, S. T. Kuna, P. K. Schweitzer, C. Kushida, P. E. Peppard, H. B. D. Sorensen, P. Jennum, and E. Mignot, "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy", *Nat. Commun.*, vol. 9, p. 5229, 2018. DOI: 10.1038/s41467-018-07229-3[2].

3.1 RESEARCH BACKGROUND

Sleep staging is important to the analysis of human sleep with about 845 000 sleep studies performed in 2014 in the US alone [75]. A standard clinical sleep study consists of a full-night PSG comprising EEG, EOG, EMG, ECG, respiratory inductance plethysmography, oronasal thermal flow, nasal pressure, and blood oxygen saturation recordings. These studies are evaluated by experts for the presence of events of clinical relevance, as determined by standards created by the AASM, such as the number of blood oxygen desaturations, micro-arousals, leg movements, periods of cessated breathing, to name a few. The overall sleep architecture is captured in a visual representation called a hypnogram, which is achieved by labeling every 30 s of PSG data into one of five stages of sleep: W, REM sleep, N1, N2, and N3, and plotting as a function of time. The latter three stages are distinguished by distinct EEG amplitude and frequency distributions, the presence of specific EEG micro-events and arousability differences reflecting sleep depth.

Sleep stage scoring is summarized in key metrics, such as the percentage of TST spent in any of the five stages (%W, or wake after sleep onset (WASO); %REM; %N1; %N2; %N3), and visually in the form of a hypnogram, which shows temporal progression of sleep stages across the night as mentioned. Current clinical practice (gold standard) of sleep study analysis is manual scoring and annotation of sleep stages and sleep events based on guidelines from the AASM [76]. These guidelines, based on observations made in healthy young males almost 70 years ago, are problematic for several reasons: a) technicians will never score the same data the exact same way as another technician, or even the same way twice [24], [26], [28], [29], [77]; b) normal sleep from healthy young males may not reflect sleep patterns of patients referred to sleep clinics; and c) the 30 s epoch rule was arbitrarily based on physical limitations of recording equipment, when PSGs were recorded on paper, and may not accurately reflect the true underlying neurobiological mechanisms.

This is described in detail in Section 2.3.1

Automatic sleep stage classification has not yet seen wide-spread adoption in clinical practice despite ongoing research demonstrating feasibility and industrial interests [78]. A major issue has been a lack of available data for designing and training models. The publicly available PhysioNet Sleep-EDF and the expanded version databases [79], [80] has been used extensively for training both shallow and deep learning-based machine learning models [81]–[83], but given the small sample size and homogeneity (most papers use the same healthy 20 subjects), it is questionable how well models derived from this data generalize to unseen data, even if high classification performance is often reported [78]. Other databases which have been extensively used

[2] Creative Commons Attributes 4.0 International License: http://creativecommons.org/licenses/by/4.0/

include the St. Vincent's University Hospital and University College Dublin Sleep Apnea Database ($n = 25$) [32], [79], and the Montreal Archive of Sleep Studies (MASS, $n = 200$) [83]–[88].

The argument for using deep learning-based models to classify high-dimensional electrophysiological data, e.g. PSGs, into discrete outcomes such as sleep stages is compelling, because of their ability to capture variability in the underlying highly complex data representations, that might be missed by machine learning methods relying on manual feature engineering. In the image, speech, and natural language processing domains, the success of deep learning models using un-transformed data has been unsurpassed in the last decade, thanks largely due to the availability of ever-increasing amounts of compute resources and more significantly very large, robust and diverse datasets [33].

Recently deep learning models for automatic sleep stage classification have been developed and validated using two or more databases or cohorts [73], [89], [90], or a single large volume cohort [72], [89], [91]. The assumption has been that by incorporating multiple sources of variance in the dataset used for training (e. g. from multiple technicians, sites, recording setups, equipment, *etc.*), final models will be better at generalizing to new, unseen data. However, no study to date has investigated multiple, large-scale cohorts for automatic sleep stage classification, or how different cohorts generalize to one another.

3.1.1 *Research motivation and objectives*

Motivated by these issues, we were interested in the following research questions specifically related to research hypothesis **RH 1**:

RH 1: *Advanced biomedical signal processing and machine learning algorithms can be used for efficient, high-performing analysis of sleep studies with regards to sleep stages.*

RQ 1.1 can sleep stages be effectively and reliably classified using novel machine learning algorithms,

RQ 1.2 in cases of multiple available data sources, is it better to have more volume or more diverse data,

RQ 1.3 how can we guarantee that such a system is stable with respect to the impact of sleep disorders.

Derived from the research hypothesis and associated questions, the following research objectives were formulated:

(i) a single model should classify sleep stages and assign probabilities to each sleep stage to allow for stage mixing;

(ii) the model should be tested using as diverse data as possible.

The following sections describe the steps taken to complete the research objectives and answer the research questions.

3.2 PAPER I: DEEP RESIDUAL NETWORKS FOR AUTOMATIC SLEEP STAGE CLASSIFICATION OF RAW POLYSOMNOGRAPHIC WAVEFORMS

ABSTRACT We have developed an automatic sleep stage classification algorithm based on deep residual neural networks and raw polysomnogram signals. Briefly, the raw data is passed through 50 convolutional layers before subsequent classification into one of five sleep stages. Three model configurations were trained on 1850 polysomnogram recordings and subsequently tested on 230 independent recordings. Our best performing model yielded an accuracy of 84.1% and a Cohen's kappa of 0.746, improving on previous reported results by other groups also using only raw polysomnogram data. Most errors were made on non-REM stage 1 and 3 decisions, errors likely resulting from the definition of these stages. Further testing on independent cohorts is needed to verify performance for clinical use.

3.2.1 *Methods*

3.2.1.1 *Data*

A database containing 2310 recordings extracted from the Wisconsin Sleep Cohort (WSC) was used in this study. Specific acquisition details concerning the PSGs are described in [92], [93]. The entire set of PSG studies was randomly split into training (train), validation (eval), and testing (test) subgroups in an 8:1:1 ratio. Detailed demographic information as well as relevant PSG variables for all three subgroups are provided in Table 3.1 including AHI and time spent in each sleep stage based on manual scoring.

3.2.1.2 *Data processing pipeline*

Central and occipital EEG from the right hemisphere, left and right EOG, and chin EMG channels were extracted from each PSG study. To accommodate different equipment setups used for recording studies, each channel was upsampled to 200 Hz. Following resampling, signals were filtered using zero-phase Butterworth filters with frequency ranges recommended by the AASM2016 [94]. Since dynamic ranges vary considerably across channels, each signal was soft-normalized using the 5th and 95th quantiles, such that

$$\mathbf{x}_{\text{norm}} = 2\,\frac{\mathbf{x} - Q_{0.05}(\mathbf{x})}{Q_{0.95}(\mathbf{x}) - Q_{0.05}(\mathbf{x})} - 1, \tag{3.1}$$

where $\mathbf{x}_{\text{norm}}$ denotes the normalized version of the signal $\mathbf{x}$, and $Q_{0.05}(\mathbf{x})$ and $Q_{0.95}(\mathbf{x})$ denotes the 5th and 95th percentile, respectively. Doubling and subtracting by one rescales $Q_{0.05}(\mathbf{x})$ and $Q_{0.95}(\mathbf{x})$ to -1 and 1, respectively.

Finally, each signal was segmented into 30 s epochs corresponding to AASM2016 criteria [94], resulting in a tensor $\mathbf{X}$ with elements

We introduce a singleton dimension, as the `tf.layers.conv1d` implementation in TensorFlow reshapes the input argument to match `tf.layers.conv2d`.

$$(x_{n,c,\cdot,t}) \in \mathbb{R}^{N\times C\times 1\times T}, \tag{3.2}$$

with $N = 16$, $C = 5$, and $T = 6000$ being batch size, number of signals, and number of timesteps for one epoch, respectively.

Table 3.1: WSC demographics for each subgroup.

	Train	**Eval**	**Test**	***p*-value**
n (male)	1850 (1010)	230 (112)	230 (120)	0.210
Age, years	59.2 ± 8.4	59.9 ± 8.5	60.4 ± 8.2	0.092
BMI, $\mathrm{kg\,m^{-2}}$	31.7 ± 7.2	31.0 ± 6.9	32.2 ± 7.7	0.203
AHI, $\mathrm{h^{-1}}$	12.6 ± 15.6	11.5 ± 14.9	12.4 ± 16.2	0.600
TST, h	7.4 ± 0.8	7.4 ± 0.7	7.4 ± 0.8	0.947
W, %	18.5 ± 11.3	17.2 ± 11.1	19.6 ± 11.8	0.071
N1, %	8.2 ± 4.5	8.8 ± 5.6	8.9 ± 5.1	**0.038**
N2, %	54.2 ± 10.3	54.0 ± 10.9	52.4 ± 11.0	**0.048**
N3, %	5.8 ± 6.4	6.4 ± 7.0	6.0 ± 7.0	0.433
REM, %	13.3 ± 5.9	13.7 ± 5.8	13.2 ± 5.7	0.635

Values are shown as mean ± standard deviation across subjects. Significant *p*-values highlighting differences between subsets are highlighted in bold as tested with χ^2 test (population proportions) and ANOVA (rest). WSC: Wisconsin Sleep Cohort; BMI: body-mass index; AHI: apnea-hypopnea index; TST: total sleep time; W: wakefulness; N1: non-rapid eye movement stage 1; N2: non-rapid eye movement stage 2; N3: non-rapid eye movement stage 3; REM: rapid eye movement.

3.2.1.3 *Deep residual network model*

We applied a deep learning model inspired by the residual network models proposed in [95], [96]. These types of models employ residual skip connections between layers in order to maintain a proper gradient backpropagation through the network. This feature allows for extremely deep network structures, and a specific variant of this model with 152 layers came in 1st place in the ILSVRC '15 image classification competition [95].

This is also known as the vanishing gradient problem and is especially problematic in very deep networks and RNNs.

network architecture The residual network model is illustrated in Figure 3.1. Briefly, the bulk network comprised 50 convolutional (conv) and dense layers arranged in four block layers of four bottlenecked residual blocks each.

A single bottleneck residual block contains three triplets of a batch normalization layer, a rectified linear unit (ReLU) activation layer, and a conv layer. This pre-activation configuration has shown benefits with regards to trainability and generalization compared to vanilla residual blocks [96]. Projection shortcuts were used between the first ReLU and conv layers to the output of the last conv layer. Kernel sizes were set to 1×1 for the first and third conv layers, and 1×3 for the second conv layer. The number of output filters for each residual block was $l \times f$ with l being the block layer index and $f = 16$, resulting in a total of 256 filters after the final conv layer.

Prior to the bottleneck blocks, the input tensor $\mathbf{X}$ was passed through an initial conv layer consisting of 64 1×16 filters, and then through a maximum pooling (max pool) layer with a 1×2 kernel and stride size, effectively reducing the time-resolution by a factor of 2. This max pool operation was implemented in the beginning of each block layer.

The output tensor from the block layers was subsequently passed to a final batch normalization and ReLU activation layer, followed by a mean pooling

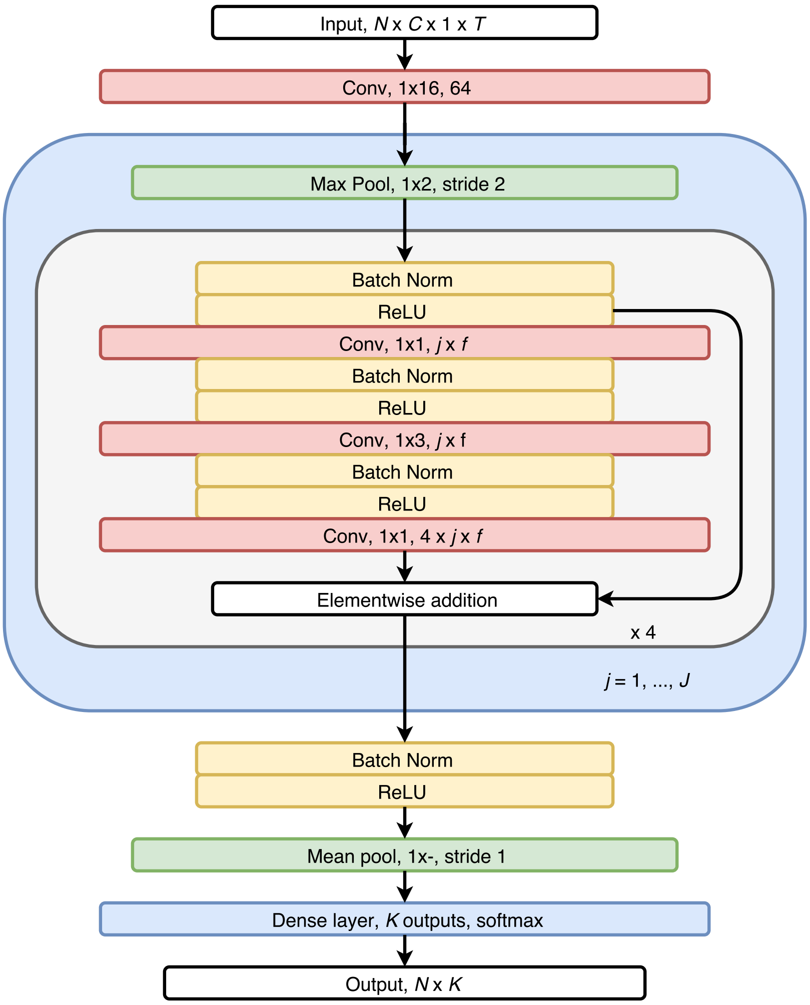


Figure 3.1: MASSC network architecture. The input tensor containing EEG, EOG, and EMG has shape $(N, C, 1, T)$, where N, $C = 5$, $T = 6000$ correspond to the batch size, number of signals, and length of each 30 s epoch, respectively. The output tensor has shape $N \times K$ with $K = 5$ sleep stages, while $L = 4$, and $f = 16$ is the number of block layers and base number of filters.

layer to reduce the tensor to $\mathbf{X} = (x_{nk}) \in \mathbb{R}^{N \times 256}$. Finally, a fully connected layer with $K = 5$ output units corresponding to the sleep stages resulted in the following output tensor

$$\mathbf{P} = (p_{nk}) \in \mathbb{R}^{N \times K}, \quad p_{nk} = \frac{\exp(z_{nk})}{\sum_k^K \exp(z_{nk})} \tag{3.3}$$

with p_{nk} containing the softmax activations of the output units z_{nk} from the fully connected layer for the nth subject and the kth sleep stage. The predicted class for the nth subject can then be calculated as

$$\hat{y}_n = \arg\max_k p_{nk}. \tag{3.4}$$

TRAINING SETUP The optimization problem was constructed using cross entropy loss across K classes and N epochs as objective function, such that

$$\mathcal{L}(\mathbf{p}_n | \mathbf{y}_n, \theta_w) = -\sum_{k=1}^{K} y_{nk} \log p_{nk}, \tag{3.5}$$

is the calculated cross entropy loss for epoch n given predicted class probabilities $\mathbf{p}_n$, true class labels $\mathbf{y}_n$, and the set of current weights θ_w. Then, the average cost across a batch of data is

$$\mathcal{C}(\mathbf{P}|\,\mathbf{Y}, \theta_w) = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(\mathbf{p}_n|\mathbf{y}_n, \theta_w). \tag{3.6}$$

The cost function was optimized using the Adam optimization algorithm with default hyperparameters [97]. Weights were initialized using variance scaling [98], and we applied weight decay during training with a decay factor of $\lambda = 10^{-4}$. The initial learning rate was set to $\alpha = 10^{-3}$ and was multiplied by 0.1 every 50 000 steps.

In order to investigate the effect of the imbalanced data on the network performance, we trained the following three different configurations. First, we defined a *baseline* configuration as described in the previous sections. The second was a *weighted* configuration, where the cost function in Equation (3.6) was replaced with an average weighted by the inverse frequency for the correct class, such that

$$\mathcal{C}(\hat{\mathbf{Y}}|\,\mathbf{Y}, \theta_w) = \frac{\sum_n^N \omega_n(\mathbf{y}_n)\mathcal{L}(\hat{\mathbf{y}}_n|\mathbf{y}_n, \theta_w)}{\sum_n^N \omega_n(\mathbf{y}_n)}, \tag{3.7}$$

where $\omega_n(\mathbf{y}_n)$ is the inverse frequency for the correct class for the nth subject in the current batch. Finally, a *balanced* configuration was tested, in which we performed resampling of the training dataset in order to balance classes. We oversampled the N1, N3, and REM classes with replacement, while undersampling the N2 class in order to have approximately equal fractions of each class in total.

3.2.1.4 *Performance metrics*

Individual precision, recall and F1 scores (Pr, Re, F1) were calculated for each sleep stage and subsequently aggregated for each recording by stage frequency weighting, such that

$$\mathrm{Pr}_{nk} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{Pr}_n = \frac{\sum_k \beta_{nk}\mathrm{Pr}_{nk}}{\sum_k \beta_{nk}} \tag{3.8}$$

$$\mathrm{Re}_{nk} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad \mathrm{Re}_n = \frac{\sum_k \beta_{nk}\mathrm{Re}_{nk}}{\sum_k \beta_{nk}} \tag{3.9}$$

$$\mathrm{F1}_{nk} = 2 \cdot \frac{\mathrm{Pr}_{nk} \cdot \mathrm{Re}_{nk}}{\mathrm{Pr}_{nk} + \mathrm{Re}_{nk}}, \quad \mathrm{F1}_n = \frac{\sum_k \beta_{nk}\mathrm{F1}_{nk}}{\sum_k \beta_{nk}}, \tag{3.10}$$

where β_{nk} is the frequency of stage k for recording n, and TP, FP and FN are true positives, false positive, and false negatives, respectively. Overall accuracy (Acc) and Cohen's kappa (κ) were also calculated for each recording. All metrics were summarized by mean and standard deviations.

3.2.1.5 *Statistical tests*

Demographic and PSG variables were tested with analysis of variances (ANOVAs) after establishing normality, while gender was tested with a χ^2 test. Significance was set at $\alpha = 0.05$.

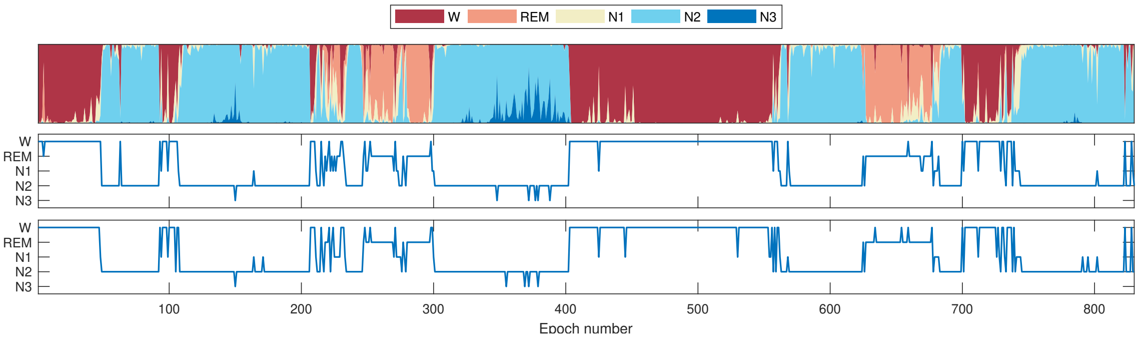


Figure 3.2: Top: hypnodensity graph of per-epoch probability distributions, middle: automatically scored hypnogram by applying Equation (3.4). Bottom: manually scored hypnogram. Note the intrusions of N3 into N2 around epoch 150 and 370, and N1 into W around 420.

Table 3.2: Averaged performance metrics for configurations.

		Baseline	Weighted	Balanced
Train	Acc, %	**86.1** ± **5.5**	79.4 ± 7.1	80.4 ± 7.3
	κ, %	**77.1** ± **8.6**	69.5 ± 9.7	70.7 ± 9.8
	Pr, %	87.1 ± 4.9	88.7 ± 4.1	**88.9** ± **4.0**
	Re, %	**86.1** ± **5.5**	79.4 ± 7.1	80.4 ± 7.3
	F1, %	**85.3** ± **6.1**	81.8 ± 6.6	82.6 ± 6.9
Eval	Acc, %	**85.0** ± **6.1**	78.4 ± 7.3	79.7 ± 7.4
	κ, %	**75.4** ± **9.5**	68.1 ± 10.5	69.7 ± 10.0
	Pr, %	86.3 ± 5.3	87.8 ± 4.8	**88.0** ± **4.9**
	Re, %	**85.0** ± **6.1**	78.4 ± 7.3	79.7 ± 7.4
	F1, %	**84.0** ± **7.2**	80.7 ± 7.1	81.9 ± 7.1

Metrics are shown as mean ± standard deviations across each PSG. Best performing model for each metric is shown in bold.

3.2.2 *Results and discussion*

Performance metrics for the train and eval subgroups are shown in Table 3.2. Not accounting for Pr, the baseline configuration compares favorably to the weighted and balanced configurations on both subgroups with an average accuracy of 85.0 % and a Cohen's kappa of 75.4 on the eval subgroup. Since the training data is imbalanced in favor of N2, it would be fair to assume overfitting to the majority class. However, the lower spread in both precision and recall does not support this.

Evaluating the baseline model on the test subgroup, only a slight drop in accuracy and κ is observed, indicating that the model generalizes well, see Table 3.3 and Table 3.4. The lowest sensitivity is obtained for N1 and N3, which is in accordance with clinical experience reported in the literature [24], [26], [28], [77]. N1 is a transitional stage between wakefulness, drowsiness and sleep often containing beta and alpha activity in epochs of low interscorer agreement, which explains the low predictive power in the confusion matrix. The sleep continuum is also apparent in Figure 3.2 which shows the manually and automatically scored hypnograms in the middle and bottom traces, and

Table 3.3: Aggregated confusion matrix and stage-specific performance metrics in test subgroup.

		Automatic							
	W	W	N2	N3	REM	Pr, %	Re, %	F1, %	
Manual	W	37980	1322	852	2	327	84.3	93.8	88.8
	W	3922	8784	3545	0	2193	51.9	47.6	49.7
	N2	1756	5136	99564	1091	991	88.6	91.7	90.2
	N3	18	1	7932	4063	14	78.8	33.8	47.3
	REM	1361	1680	465	0	23931	87.2	87.2	87.2

W: wakefulness; N1: non-rapid eye movement stage 1; N2: non-rapid eye movement stage 2; N3: non-rapid eye movement stage 3; REM: rapid eye movement.

Table 3.4: Performance across recordings in test subgroup.

Accuracy	κ	**Pr**	**Re**	**F1**
84.1 ± 6.9	0.746 ± 0.099	85.7 ± 6.1	84.1 ± 6.9	83.1 ± 7.6

Values are shown as mean ± standard deviation across PSG. Pr: precision; Re: recall.

the hypnodensity graph in the top trace for a representative subject in the test subgroup. The hypnodensity is a probabilistic representation of the hypnogram, which has found use in the detection of narcolepsy and Parkinson's disease [73], [99], [100].

Our baseline model attains favorable performance when comparing to the results reported for the raw waveform CNN model in [91] with both higher accuracy and Cohen's kappa. However, it should be stressed that [91] only used EEG channels from 9000 recordings, while our model uses both EEG, EOG and EMG data, but only from 1850 recordings. Furthermore, our baseline model performs only slightly worse compared to the best-performing model in [91] without using any memory networks. This indicates a performance gain by adding recurrent networks, such as long short-term memory cells, to our network.

A possible limiting factor to our model is the filter kernels. The small filter sizes in block layers might not be able to accurately capture the physiological dynamics, but there are indications that many, smaller kernels are preferable to fewer, larger kernels when comparing model complexity versus computational costs [101].

Future work will include adding more data to balance classes, and adding long short-term memory cells to the network in order to model temporal dynamics between epochs.

3.3 Paper II: Automatic sleep stage classification with deep residual networks in a mixed-cohort setting

Study objectives: Sleep stage scoring is performed manually by sleep experts and is prone to subjective interpretation of scoring rules with low intra- and interscorer reliability. Many automatic systems rely on few small-scale databases for developing models, and generalizability to new datasets is thus unknown. We investigated a novel deep neural network to assess the generalizability of several large-scale cohorts.

Methods: A deep neural network model was developed using 15 684 polysomnography studies from five different cohorts. We applied four different scenarios: 1) impact of varying time-scales in the model; 2) performance of a single cohort on other cohorts of smaller, greater or equal size relative to the performance of other cohorts on a single cohort; 3) varying the fraction of mixed-cohort training data compared to using single-origin data; and 4) comparing models trained on combinations of data from 2, 3, and 4 cohorts.

Results: Overall classification accuracy improved with increasing fractions of training data (0.25%: 0.782 ± 0.097, 95 % CI: $[0.777 - 0.787]$; 100%: 0.869 ± 0.064, 95 % CI: $[0.864 - 0.872]$), and with increasing number of data sources (2: 0.788 ± 0.102, 95 % CI: $[0.787 - 0.790]$; 3: 0.808 ± 0.092, 95 % CI: $[0.807 - 0.810]$; 4: 0.821 ± 0.085, 95 % CI: $[0.819 - 0.823]$). Different cohorts show varying levels of generalization to other cohorts.

Conclusions: Automatic sleep stage scoring systems based on deep learning algorithms should consider as much data as possible from as many sources available to ensure proper generalization. Public datasets for benchmarking should be made available for future research.

3.3.1 *Cohort descriptions*

To investigate and conclude on generalizability of any machine learning or sleep stage classification model, multiple heterogenous datasets must be used for training, validation and testing purposes. In this work, we collected datasets from five different sources, each dataset containing a diverse collection of subjects presenting with multiple disease phenotypes. Details of the separate cohorts are shown in Table 3.5 along with reported p-values highlighting cohort differences. Each cohort was split into a training, validation and testing subset in proportions of 87.5 %, 2.5 %, and 10 %, respectively, using random sampling without replacement among unique subjects, so that no subject is shared between subsets. With these percentages, we maximize the number of PSGs available for training, while still reserving enough PSGs for validation and testing. Collecting all the separate subsets across cohorts forms a training, validation, and testing partition, containing the respective subsets from all five cohorts.

3.3.1.1 *Institute of Systems and Robotics, University of Coimbra Sleep Cohort (ISRUC)*

This cohort contains 126 recordings from 118 unique subjects recorded at the Sleep Medicine Centre of the Hospital of Coimbra University, Portugal, in the period 2009–2013 [102]. The cohort comprises three subgroups: subgroup I contains 100 PSGs of subjects with diagnosed sleep disorders, generally sleep apnea; subgroup II contains 16 recordings of eight subjects most of which are also diagnosed with sleep apnea; and subgroup III contains recordings from 10 subjects with no diagnosed sleep disorders. All PSGs were recorded with the same recording hardware and software and each was scored by two technicians for sleep stages and sleep events according to the AASM guidelines. ISRUC is a freely accessible resource and all data and PSG files can be located at `https://sleeptight.isr.uc.pt/ISRUC_Sleep/`.

3.3.1.2 *Osteoporotic Fractures in Men Sleep Study (MrOS)*

The MrOS Sleep Study is part of the larger Osteoporotic Fractures in Men Study, which aims to understand the relationships between sleep disorders, fractures, and vascular diseases in community-dwelling men [103]–[105]. It consists of 2907 in-home PSG recordings with an additional 1026 follow-up PSG studies from subjects recruited from six different clinical centers in the USA. Each recording was annotated by an expert technician according to Rechtschaffen and Kales (R&K) criteria for sleep staging [106]. For compatibility with AASM guidelines, we combined stages labeled S3 and S4 into N3. All data were accessed from the National Sleep Research Resource (NSRR) repository [107], [108].

3.3.1.3 *SHHS*

The SHHS is a large, multi-center study on cardiovascular outcomes related to sleep disorders with a specific focus on sleep-disordered breathing [109], [110]. The cohort consists of 6441 subjects above 40 years old recruited between 1995 and 1998 undergoing in-home PSG (SHHS Visit 1) with subsequent follow-up PSG between 2001 and 2003 in 3295 subjects (SHHS Visit 2). PSG recordings were annotated for sleep stages by trained and certified technicians according to R&K rules. From the original cohort we extracted 5793 PSGs and annotations from Visit 1, and 2651 from Visit 2. We aggregated S3 and S4 stages into N3 similar to MrOS. All data were accessed from NSRR repository.

3.3.1.4 *WSC*

WSC is a population-based study of sleep-disordered breathing in government workers in Wisconsin, USA, that was initiated in 1988 [92], [93]. In this work, we used 2412 PSGs from 1091 unique subjects in the WSC sample scored by expert technicians according to R&K rules with subsequent merging of S3 and S4 into N3.

3.3.1.5 *Stanford Sleep Cohort (SSC)*

PSGs from this cohort originate from patients referred for sleep disorders evaluation and recorded at the Stanford Sleep Clinic since 1999. The specific sample used in this study represents a small subset ($n = 772$) of the whole cohort, which was selected and described in detail in previous studies scored according to R&K or AASM guidelines according to prevailing standard at the time of evaluation [111], [112].

Table 3.5: Cohort demographics

	ISRUC	**MrOS**	**SHHS**	**SSC**	**WSC**	***p*-value**
N (female)	126 (50)	3932 (0)	8444 (4458)	767 (319)	2401 (1103)	—
Age, years	49.8±15.9 [20.0–85.0]	77.6±5.6 [67.0–90.0]	64.5±11.2 [39.0–90.0]	45.7±14.5 [13.0–104.8]	59.7±8.4 [37.2–82.3]	<0.0001
BMI, kg/m2	—	27.1±3.8 [16.0–47.0]	28.2±5.1 [18.0–50.0]	27.2±6.5 [9.8–78.7]	31.6±7.2 [17.5–70.6]	<0.0001
TST, min	350.0±67.3 [87.5–479.0]	352.1±71.9 [39.0–626.0]	374.1±69.4 [68.0–605.0]	361.0±83.5 [0.0–661.0]	364.1±63.6 [19.5–575.0]	<0.0001
SL, min	17.7±20.5 [0.0–144.5]	24.7±26.9 [1.0–402.0]	24.2±25.7 [0.0–349.0]	93.5±58.9 [0.5–404.0]	33.2±21.4 [0.5–333.0]	<0.0001
REML, min	125.6±61.4 [7.0–323.0]	104.8±75.1 [0.0–590.0]	91.7±58.8 [0.0–471.0]	140.9±88.0 [0.0–464.0]	128.3±76.0 [3.5–514.0]	<0.0001
WASO, min	76.2±49.8 [7.5–251.0]	117.5±67.6 [4.0–487.0]	80.2±54.7 [2.0–378.0]	79.5±55.0 [3.5–367.0]	73.6±45.9 [3.0–325.0]	<0.0001
SE, %	78.8±14.1 [19.5–98.3]	75.5±12.4 [12.0–99.0]	80.5±11.0 [11.3–99.0]	77.4±14.8 [0.0–98.0]	77.1±11.2 [4.1–95.6]	<0.0001
N1, %	13.3±5.8 [1.8–33.1]	8.3±6.4 [0.0–70.0]	5.5±4.0 [0.0–39.1]	11.7±10.2 [0.0–92.0]	10.8±6.9 [1.0–88.4]	<0.0001
N2, %	31.9±10.3 [4.4–89.3]	62.5±10.0 [21.0–95.0]	56.9±11.5 [10.9–100.0]	62.8±24.9 [0.0–636.0]	66.0±9.4 [9.1–93.3]	<0.0001
N3, %	19.6±8.0 [0.0–41.1]	36.0±31.8 [0.0–259.0]	17.5±11.6 [0.0–70.1]	9.0±9.3 [0.0–73.0]	7.2±7.8 [0.0–47.5]	<0.0001
REM, %	13.3±6.3 [0.0–37.8]	19.3±6.8 [0.0–44.0]	20.1±6.3 [0.0–48.0]	16.3±7.2 [0.0–40.0]	16.0±6.2 [0.0–38.2]	<0.0001
ArI, /h	20.2±10.0 [2.1–72.0]	23.7±12.1 [1.0–105.0]	18.9±10.5 [0.0–110.4]	125.0±124.2 [1.0–729.0]	—	<0.0001
AHI, /h	13.1±13.2 [0.0–82.2]	13.7±14.6 [0.0–89.0]	18.1±16.2 [0.0–161.8]	13.5±19.2 [0.0–98.6]	7.0±9.4 [0.0–72.6]	<0.0001
PLMI, /h	8.0±27.4 [0.0–292.8]	35.7±37.5 [0.0–233.0]	—	7.0±18.1 [0.0–139.9]	—	<0.0001

BMI: body-mass index; TST: total sleep time; SL: sleep latency; REML: REM sleep latency; WASO: wake after sleep onset; SE: sleep efficiency; N1: non-rapid eye movement stage 1; N2: non-rapid eye movement stage 2; N3: non-rapid eye movement stage 3; REM: rapid eye movement; ArI: arousal index; AHI: apnea-hypopnea index; PLMI: periodic leg movement index; ISRUC: Institute of Systems and Robotics, University of Coimbra Sleep Cohort; MrOS: Osteoporotic Fractures in Men Sleep Study; SHHS: Sleep Heart Health Study; SSC: Stanford Sleep Cohort; WSC: Wisconsin Sleep Cohort;

3.3.2 *Methods*

3.3.2.1 *Data pipeline*

Electrophysiological signals corresponding to the minimum acceptable montage for sleep staging available across all cohorts were extracted for each PSG. These included a central EEG (either C3 or C4 referenced to the contra-lateral mastoid), left and right EOG referenced to the contra-lateral mastoid, and a single submentalis EMG. The choice between C3 and C4 was determined based on the lowest total signal energy across the entire duration of the PSG to avoid excessive signal popping. Other methods to determine appropriate channels include algorithms based on shortest Mahalanobis distance to an already determined reference distribution [73], but was not investigated in this study. All signals were resampled to $f_s = 128\,\text{Hz}$ using a polyphase filtering procedure irrespective of original sampling frequency, and subsequently filtered using a zero-phase approach with 4th order Butterworth IIR filters (0.5 Hz to 35 Hz band-pass for EEG and EOG; 10 Hz high-pass for EMG) in accordance with AASM2020 filter specifications [43]. Each signal was normalized to zero mean and unit variance to accommodate differences in recording equipment and baselines, and to compress the dynamic range into something easily trainable for the neural network architecture. We denote by C the number of input signals supplied to the neural network, where in this case $C = 4$.

3.3.2.2 *Machine learning problem*

We designate by $\mathcal{X} \in \mathbb{R}^{C \times T}$ the set of 30 s input data segments with S input channels and segment length T, and the corresponding sleep stage classifications by $\mathcal{Y} = \{y \in \mathbb{R}^K_+ \mid \sum_i y_i = 1\}$, where $K = 5$ corresponds to the five sleep stages. Thus, y is a probability simplex, which maps to the ordered set $\mathcal{S} = \{\text{W}, \text{N1}, \text{N2}, \text{N3}, \text{R}\}$ by the argmax function such that $\arg\max y : \mathcal{Y} \to \mathcal{S}$. Furthermore, as we are potentially interested in classifying multiple sleep stages at once, we extend the problem of classifying a single sleep stage given $x \in \mathcal{X}$ to a sequence-to-sequence problem, in which we desire to learn a differentiable function representation Φ, that maps a sequence of 30 s epochs $\mathbf{x} \in \mathbb{R}^{C \times \alpha T}$ to their corresponding label probabilities $\mathbf{y} \in \mathbb{R}^{K \times \alpha}$, where α is a parameter that controls the sequence length. For example, if $\alpha = 8$, the sequence $\mathbf{x}$ contains 4 min of successive PSG data described by 8 epochs of length 30 s. Furthermore, we denote by $[\![a, b]\!]$ the set of integers from a to b, i.e. $[\![a, b]\!] = \{n \in \mathbb{N} \mid a \leqslant n \leqslant b\}$, and by $[\![N]\!]$ the shorthand form of $[\![1, N]\!]$.

3.3.2.3 *Network architecture*

As the representation of Φ, we adapted and extended a previously published neural network architecture for automatic sleep stage classification, which was based on a variant of the ResNet-50 architecture commonly used for two-dimensional image classification tasks, but adapted and re-trained from scratch for the specific use-case of one-dimensional, time-dependent signals in the PSG [72]. This network has the advantage that it does not require any manual feature engineering and extraction compared to previous state of the art sleep stage classification models [73]. An overview of the proposed network architecture is provided graphically in Figure 3.3 and Table 3.6. Briefly, the architecture consists of four modules:

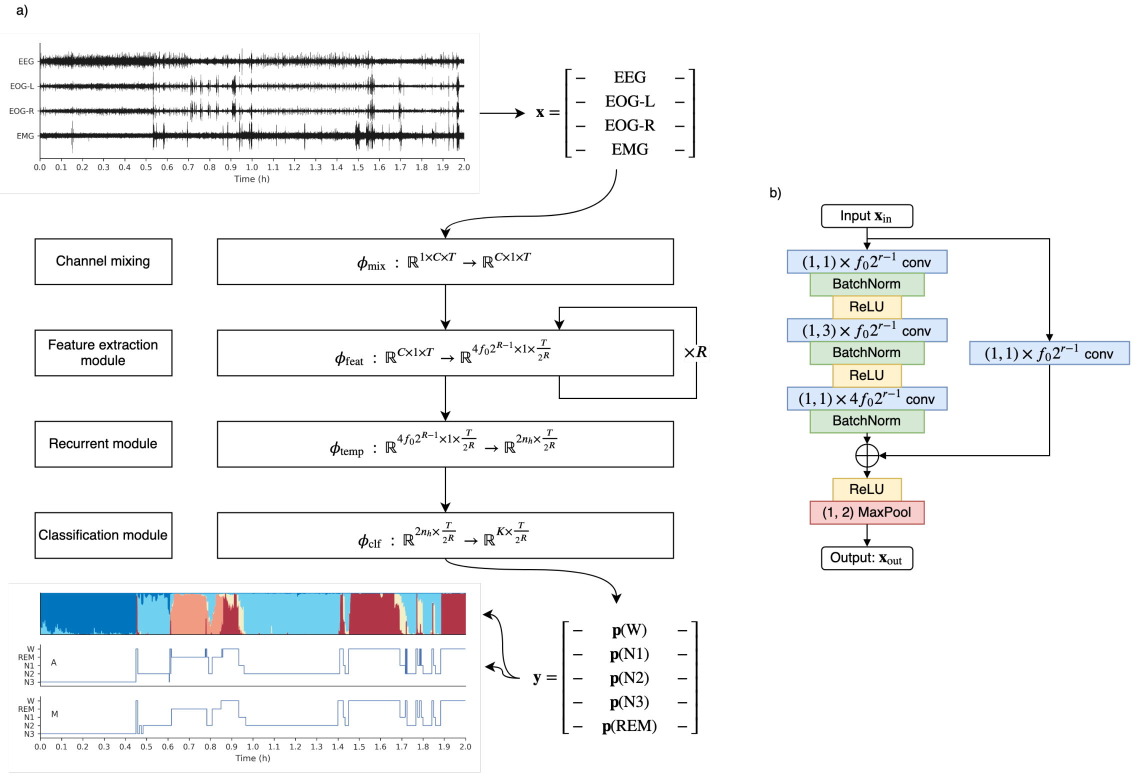

a)
EEG
EOG-L
EOG-R
EMG
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
Time (h)
$\mathbf{x} = \begin{bmatrix} - & \text{EEG} & - \\ - & \text{EOG-L} & - \\ - & \text{EOG-R} & - \\ - & \text{EMG} & - \end{bmatrix}$
Channel mixing
$\phi_{\text{mix}} : \mathbb{R}^{1\times C\times T} \to \mathbb{R}^{C\times 1\times T}$
Feature extraction module
$\phi_{\text{feat}} : \mathbb{R}^{C\times 1\times T} \to \mathbb{R}^{4f_0 2^{R-1}\times 1\times \frac{T}{2^R}}$
$\times R$
Recurrent module
$\phi_{\text{temp}} : \mathbb{R}^{4f_0 2^{R-1}\times 1\times \frac{T}{2^R}} \to \mathbb{R}^{2n_h\times \frac{T}{2^R}}$
Classification module
$\phi_{\text{clf}} : \mathbb{R}^{2n_h\times \frac{T}{2^R}} \to \mathbb{R}^{K\times \frac{T}{2^R}}$
W
REM
N1
N2
N3
A
M
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
Time (h)
$\mathbf{y} = \begin{bmatrix} - & \mathbf{p}(\text{W}) & - \\ - & \mathbf{p}(\text{N1}) & - \\ - & \mathbf{p}(\text{N2}) & - \\ - & \mathbf{p}(\text{N3}) & - \\ - & \mathbf{p}(\text{REM}) & - \end{bmatrix}$
b)
Input $\mathbf{x}_{\text{in}}$
$(1,1)\times f_0 2^{r-1}$ conv
BatchNorm
ReLU
$(1,3)\times f_0 2^{r-1}$ conv
BatchNorm
ReLU
$(1,1)\times 4 f_0 2^{r-1}$ conv
BatchNorm
$(1,1)\times f_0 2^{r-1}$ conv
ReLU
(1, 2) MaxPool
Output: $\mathbf{x}_{\text{out}}$

Figure 3.3: Model overview. a) The input is a sequence of data $\mathbf{x}$ containing raw signal data from EEG, left/right EOG, and EMG channels, which is supplied to the network modules in sequence. The feature extraction module consists of R repeated blocks of residual units, see b) panel to the right. The output of the model is a matrix $\mathbf{y}$ containing class probabilities for each sleep stage for each time step, which can be visualized either directly as a hypnodensity, or by $\arg\max \mathbf{y}$ as a hypnogram. The **A** and **M** labels in the hypnogram plots corresponds to automatic and manually scored hypnograms. b) Schematic of a single residual block in the feature extraction module. Convolutional layers are described by the kernel size × number of filters using a stride value of 1. Shortcut uses 1×1 convolutions with added zero-padding to maintain temporal dimension. Conv, convolutional layer; BatchNorm, batch normalization; ReLU: rectified linear unit; f_0, base number of filters ($f_0 = 4$).

1. an initial mixing module

$$\varphi_{\text{mix}} : \mathbb{R}^{1\times C\times T} \to \mathbb{R}^{C\times 1\times T}, \tag{3.11}$$

2. a feature extraction module

$$\varphi_{\text{feat}} : \mathbb{R}^{C\times 1\times T} \to \mathbb{R}^{f_0 2^{R+1}\times 1\times T/2^R}, \tag{3.12}$$

3. a temporal processing module

$$\varphi_{\text{temp}} : \mathbb{R}^{f_0 2^{R+1}\times 1\times T/2^R} \to \mathbb{R}^{2n_h\times T/2^R}, \text{ and} \tag{3.13}$$

4. a classification module

$$\varphi_{\text{clf}} : \mathbb{R}^{2n_h\times T/2^R} \to \mathbb{R}^{K\times T/2^R}. \tag{3.14}$$

Thus, we obtain a differentiable representation of the function Φ as

$$\begin{aligned} &\Phi : \mathbb{R}^{C\times K} \to \mathbb{R}^{K\times T/2^R} \\ &\Phi(\mathbf{x}) = \varphi_{\text{clf}}(\varphi_{\text{temp}}(\varphi_{\text{feat}}(\varphi_{\text{mix}}(\mathbf{x})))). \end{aligned} \tag{3.15}$$

The output of this function is the matrix $\mathbf{y} \in \mathbb{R}^{K\times T/2^R}$ containing sleep stage probabilities in the sequence of PSG data evaluated every second.

Table 3.6: Overview of model architecture.

Module	Type	Filters	Kernel	Stride	Activation	Output size
$\mathbf{x}$	Input	—	—	—	—	$1 \times C \times T$
φ_{mix}	2D conv.	C	$(1, C)$	1	BN+ReLU	$C \times 1 \times T$
$\varphi_{\text{feat}}^{(r)}$	Res. block†	$f_0 2^{r-1}$	$(1, 1)$	$(1, 1)$	BN+ReLU	$f_0 2^{r-1} \times 1 \times \frac{T}{2^{r-1}}$
$r \in [\![R]\!]$		$f_0 2^{r-1}$	$(1, 3)$	$(1, 2)$	BN+ReLU	$f_0 2^{r-1} \times 1 \times \frac{T}{2^r}$
		$4 f_0 2^{r-1}$	$(1, 1)$	$(1, 1)$	BN+ReLU	$f_0 2^r \times 1 \times \frac{T}{2^r}$
φ_{temp}	bGRU	n_h	—	—	—	$2n_h \times \frac{T}{2^R}$
φ_{clf}	1D conv.	K	$2n_h$	1	Softmax	$K \times \frac{T}{2^R}$

Kernel sizes correspond to the first, second and third convolutional layer in each residual block. Stride counts correspond to the residual block and the subsequent maxpooling operation. Conv., convolution; Res. block, residual block; BN: batch normalization, ReLU: rectified linear unit, bGRU: bidirectional gated recurrent unit, C, number of input channels; T, length of segment in samples; f_0, base number of filters in residual blocks; R, number of residual blocks; n_h, number of hidden units in bGRU; K, number of sleep stage classes; †See Figure 3.3 for details.

MIXING MODULE The raw input data is input to this module, which encourages non-linear channel mixing similar to what has been proposed in recent literature [85], [113]–[115]. The module is realized using a single 2D convolutional operation outputting C feature maps computed using single-strided $(C \times 1)$ kernels followed by ReLU activations.

FEATURE EXTRACTION (RESIDUAL NETWORK) MODULE This is comprised of a succession of R residual blocks, which are responsible for the bulk feature extraction from the channel-mixed data, see Figure 3.3. Each residual block is realized using bottlenecks of first a 1×1 convolution to reduce the number of feature maps, then a 1×3 convolution, and lastly a 1×1 convolution to finally increase the number of feature maps. Each convolution operation is followed by a batch normalization [116] and ReLU activation except after the last convolutional layer, where shortcut projections are added before the activation [96]. This type of block structure enables the design and training of very deep networks without the risk of vanishing gradients due to the projection shortcuts [95].

TEMPORAL PROCESSING MODULE This module is realized by a bidirectional gated recurrent unit (GRU) [117] in order to accommodate temporal dependencies in the PSG. The GRU runs through the temporal dimension of the output from φ_{feat} of $T/2^R$ time steps each containing $f_0 2^R$ feature maps and outputs n_h new features in each direction for each time step. By running both forward and backward, we can accommodate that technicians base their scoring on looking backwards as well as ahead in time in each time segment (typically 30 s).

CLASSIFICATION MODULE The final module in the architecture performs actual classification based on the forward and backward features for each time step outputted from φ_{temp}. It is realized by a single convolutional operation with a subsequent softmax activation to compute a probability distribution over the K sleep stage classes, such that the probability of sleep stage i at time step n is given by

$$y_i^{(n)} = \frac{\exp a_i}{\sum_k \exp a_k}, \tag{3.16}$$

where $a_i \in \mathbf{a}$ is the activation of the last layer in the network, and $k \in [\![K]\!]$.

3.3.2.4 *Loss function specification*

The network was trained end-to-end with respect to a loss function, that takes the output probabilities from the network $y = \Phi(\mathbf{x})$ and calculates the loss as

$$\begin{aligned} \mathcal{L}(\mathbf{y}) &= -\sum_{n=1}^{30/\tau} \sum_{k=1}^{K} t_k^{(n)} \log\left(\tilde{y}_k^{(n)}\right), \\ \tilde{y}_k^{(n)} &= \frac{1}{\tau} \sum_{i=\tau(j-1)+1}^{\tau n} y_k^{(i)}, \end{aligned} \tag{3.17}$$

which is the cross-entropy between successive time-averaged classifications parameterized by the number of successive one-second predictions τ, and the ground truth labels t broadcasted to $30/\tau$ labels per 30 s segment. This way, we can acquire predictions every second, that can be combined in time at intervals given by τ.

3.3.2.5 *Experimental setups*

We set up four different experiments in this study.

1. We wished to investigate the effect of increasing the complexity of the recurrent module by varying the number of units n_h in the module φ_{temp} in the space $n_h = 2^k, k \in [\![6, 11]\!]$. We hypothesize that there exists a sweet-spot in the number of hidden units that balances computational complexity with classification performance, i.e. classifying a sequence of sleep stage labels given a corresponding sequence of outputs from φ_{feat}. The results of this experiment were furthermore used to determine parameters for models in subsequent experiments.

2. Since we have several cohorts at our disposition of both clinical and research origin, we can investigate the compatibility and inherent generalizability of the different cohorts in two ways: 1) we set aside a single cohort for testing, while we train the models on the remaining four (leave-one-cohort-out, leave-one-cohort-out (LOCO) training); and 2) we train on a single cohort, while we set aside the remaining four for testing (leave-one-cohort-in, leave-one-cohort-in (LOCI) training).

3. Generalizability can also be investigated in another way, which can answer the question of how many data sources is necessary. We trained models with all possible 2-, 3-, and 4-combinations of cohorts, i.e. one run trained on ISRUC and MrOS training data, another run with ISRUC and SHHS train data, a third with ISRUC and SSC, etc., with all runs subjected to subsequent evaluation on the test partition.

4. Previous studies have already investigated the performance of automatic sleep staging algorithms using shallow machine learning models. At the time of writing however, none have investigated the effect of

available training data for deep learning models at this magnitude (up to tens of thousands). We therefore trained models on 0.25 %, 0.5 %, 1 %, 5 %, 10 %, 25 %, 50 %, 75 % and 100 % of the data available for training. Specifically, some of these fractions of the total number of PSGs correspond roughly to the number of PSGs in the training partitions in each cohort, allowing for direct comparisons between training a model with mixed- and single-cohort training data.

Common for all experiments were the default parameter values $C = 4$, $f_s = 128\,\text{Hz}$, $T = \tau f_s$, $K = 5$, $R = 7$, and $f_0 = 4$ for the number of input channels, sampling frequency, the sequence length, the number of sleep stages, the number of consecutive residual blocks, and the base filter kernel size, respectively. All models were trained for 50 epochs (passes through the training partition) and the model with the highest Cohen's kappa value on the validation partition was subsequently selected for testing. All models were trained end-to-end with backpropagation using the Adam optimizer [97] with a learning rate of 10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$ to minimize the loss function specified by Equation (3.17). All network weights and bias terms were initialized using the uniform Glorot initialization scheme [118].

3.3.2.6 *Performance metrics and model evaluation*

For each experiment we evaluated model performance using the overall accuracy (Acc) and Cohen's κ in order to take account the possibility of chance agreement between the model and the gold standard. Given a confusion matrix $\mathbf{C}$ with element c_{ij} being the number of epochs belonging to sleep stage i but classified to be in sleep stage j, we define the overall accuracy for a given model as

$$\text{Acc} = \frac{\sum_{i=j} c_{ij}}{\sum_{i,j} c_{ij}}, \tag{3.18}$$

i. e. the sum of the trace of $\mathbf{C}$ divided by the total count. The Cohen's κ metric is defined as

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{3.19}$$

where $p_o = \text{Acc}$ is the observed agreement (i. e. accuracy) and p_e is the expected chance agreement, which can be reformulated in terms of the outer product between the row and column sums (class-specific recall and precision) of $\mathbf{C}$.

3.3.3 *Results*

In this section we report on the results of the three experiments described in Section 3.3.2.5.

3.3.3.1 *Temporal context impact on model performance*

In Figure 3.4 we show how the model performance depends on the temporal context and complexity of the temporal processing module, when evaluating the model on the validation partition. Results are further detailed in Table 3.7. Specifically, we observe a drastic change in Cohen's κ just by introducing a simple recurrent unit into the network as shown in Figure 3.4a, where Cohen's κ increases from 0.645 ± 0.126, 95 % CI: $[0.633 - 0.657]$ at $n_h = 0$

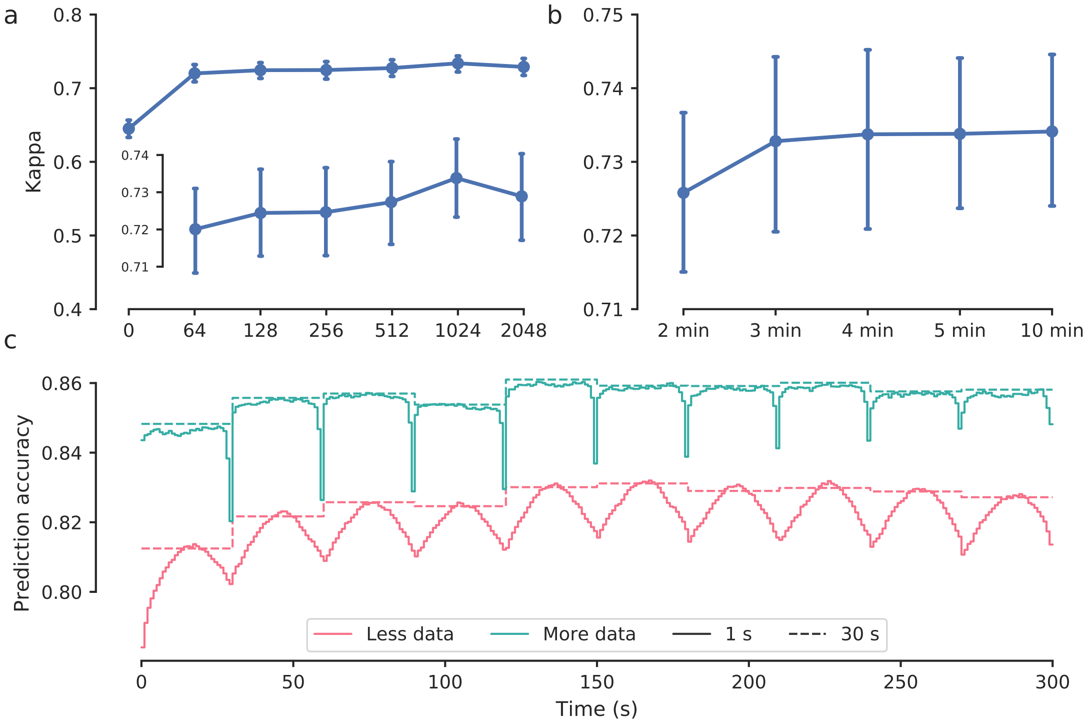


Figure 3.4: Temporal context changes model performance. a) Cohen's kappa as a function of the number of hidden units in the recurrent block. Inset shows zoom of Cohen's kappa for non-zero hidden unit values. b) Cohen's kappa as a function of sequence length. c) Prediction accuracy averaged across all 5-minute sequences in the test partition with a small and large training partition. Full lines are predictions evaluated every 1 s, while dashed lines show predictions averaged every 30 s. Values are shown for panels a), b) as mean with 95% confidence intervals.

to 0.720 ± 0.120, 95 % CI: [0.709 − 0.731] at $n_h = 64$. We did not observe any major changes when increasing the number of hidden units beyond $n_h = 64$, although we did see a maximum Cohen's κ of 0.734 ± 0.111, 95 % CI: [0.723 − 0.744] at $n_h = 1024$, which is shown in the inset in Figure 3.4a. We observed a general increase in Cohen's κ when classifying longer sequences than 2 min (0.726 ± 0.114, 95 % CI: [0.715 − 0.737]), but did not see any major differences when classifying over more than 3 min sequences (0.733 ± 0.123, 95 % CI: [0.721 − 0.7444]). Subsequent models were fixed with $n_h = 1024$ corresponding to a sequence length of 5 min.

3.3.3.2 *Model classifications converge to 30 s predictions given sufficient training data*

Furthermore, we analyzed the classification performance of the model given a specific sequence length by looking at the average prediction accuracy across all 5 min sequences in all subject PSGs in the test partition, similar to what Brink-Kjaer *et al.* has shown previously [119]. In Figure 3.4c, we show how the average classification accuracy in a 5 min sequence both depends on the amount of data and the frequency of evaluating the model output, i. e. every 1 s or across 30 s. The average classification accuracy was found to be slightly lower in the beginning of each 5 min sequence, both when training a model with less (500 training subjects) and more (75 % of total training subjects). Interestingly, when training with less data, we also observed a

Table 3.7: Temporal context impact on model performance in validation partition ($n = 426$).

	Overall accuracy				Cohen's kappa			
	Mean	**SD**	**Median**	**95% CI**	**Mean**	**SD**	**Median**	**95% CI**
Hidden units								
0	0.779	0.083	0.794	[0.771-0.787]	0.645	0.126	0.660	[0.633-0.657]
64	0.818	0.079	0.837	[0.810-0.825]	0.720	0.120	0.745	[0.709-0.731]
128	0.821	0.080	0.841	[0.813-0.829]	0.724	0.121	0.745	[0.713-0.736]
256	0.820	0.082	0.843	[0.812-0.828]	0.725	0.124	0.751	[0.713-0.736]
512	0.822	0.079	0.841	[0.815-0.830]	0.727	0.119	0.752	[0.716-0.739]
1024	0.828	0.072	0.845	[0.821-0.835]	0.734	0.111	0.758	[0.723-0.744]
2048	0.823	0.080	0.843	[0.816-0.831]	0.729	0.122	0.757	[0.717-0.740]
Sequence length								
2 min	0.821	0.075	0.840	[0.814-0.828]	0.726	0.114	0.754	[0.715-0.737]
3 min	0.826	0.080	0.845	[0.818-0.833]	0.733	0.123	0.762	[0.721-0.744]
4 min	0.828	0.079	0.849	[0.820-0.835]	0.734	0.122	0.762	[0.722-0.745]
5 min	0.828	0.072	0.845	[0.821-0.835]	0.734	0.111	0.758	[0.723-0.744]
10 min	0.829	0.075	0.848	[0.822-0.836]	0.734	0.113	0.759	[0.723-0.745]
Window length								
1 s	0.824	0.074	0.843	[0.817-0.831]	0.728	0.113	0.752	[0.717-0.738]
3 s	0.824	0.074	0.845	[0.817-0.832]	0.728	0.113	0.752	[0.717-0.739]
5 s	0.825	0.074	0.843	[0.818-0.832]	0.728	0.113	0.752	[0.717-0.739]
10 s	0.825	0.074	0.844	[0.818-0.832]	0.729	0.113	0.753	[0.718-0.739]
15 s	0.826	0.074	0.845	[0.818-0.833]	0.729	0.113	0.755	[0.719-0.740]
30 s	0.829	0.075	0.848	[0.822-0.836]	0.734	0.113	0.759	[0.723-0.745]

The hidden units variable corresponds to varying the complexity in the recurrent module by increasing the number of hidden units. Sequence length indicate the length of the sequence of 30 epochs, while window length correspond to varying the evaluation frequency. Means, standard deviations (SD) and medians are based on performance for each PSG.

lower accuracy in the beginning and end of each 30 s segment relative to the accuracy in the middle section, which was not the case when training with more data.

3.3.3.3 *Choice of cohort impacts classification performance on test set*

In Figure 3.5 we show how training on different cohorts yield differing results in subsequent testing performance, here expressed in heatmaps as both overall accuracy (Figure 3.5a), and Cohen's κ (Figure 3.5b) averaged across all $N = 1584$ subject PSGs in the test partition. The first two columns show the performance on the cohort on the x-axis, when training on the specific cohort on the y-axis. Since the training subset in ISRUC is small compared to the other cohorts, we trained the models in the left-most column with weight decay of 10^{-4} to compensate for the risk of overfitting, however, by comparing the left and middle columns, we did not observe any specific gain in classification performance by doing so. The right-most column shows the test performance for each cohort, when excluding that cohort from training. We observe a significant spread in classification accuracy across the different

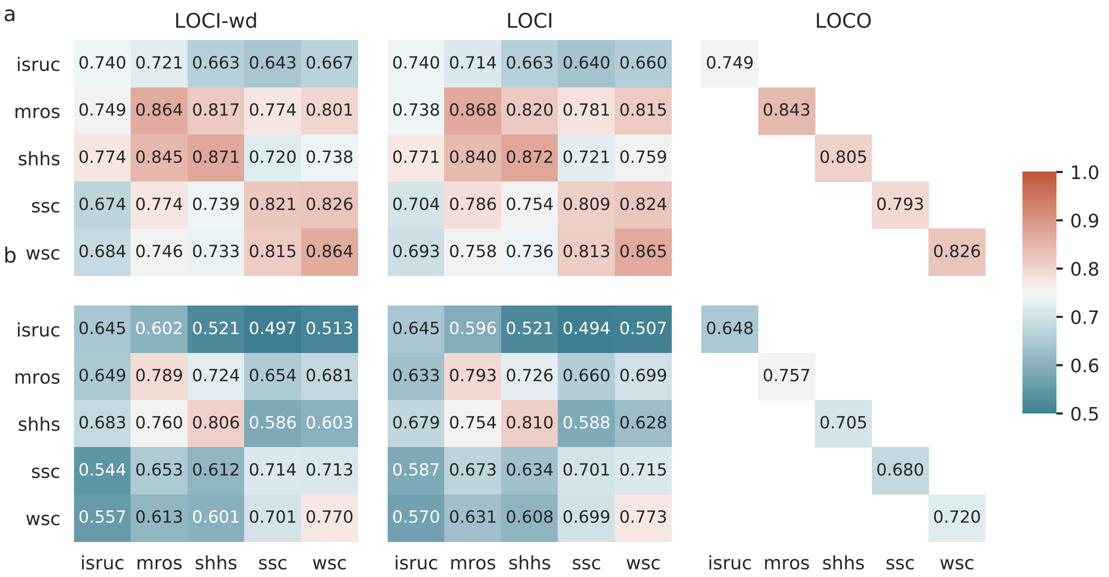


Figure 3.5: Individual cohorts influence classification performance on test partition (N = 1, 584). As an example, training on MrOS in a LOCI configuration, the performance on the test subset of WSC is 0.815. The diagonals in all three configurations shows the performance for the same subjects in the test subsets in the respective cohorts making possible direct comparisons between LOCI and LOCO. For aggregated metrics and more summary statistics, please see Table 4. LOCI: leave-one-cohort-in; LOCI-wd: LOCI with weight decay; LOCO: leave-one-cohort-out; ISRUC: Institute of Systems and Robotics, University of Coimbra Sleep Cohort; MrOS: Osteoporotic Fractures in Men Sleep Study; SHHS: Sleep Heart Health Study; SSC: Stanford Sleep Cohort; WSC: Wisconsin Sleep Cohort.

cohorts with prediction on ISRUC being poorest, while prediction on MrOS data being best. Further details can be found in Table 3.8.

3.3.3.4 *More data is good, diverse data is better*

We observed a general increase in classification performance both in terms of overall accuracy and Cohen's κ, when including more data in the model training phase in both the mixed- and single-cohort setting (Figure 3.6a, Table 3.9). Classification performance was consistently lower in the single-cohort setting compared to the corresponding mixed-cohort setting. Interestingly, we found that training a model with just 0.25 % of mixed-cohort training data still achieved an acceptable accuracy comparable to training a model with only SHHS data, while using all available training data increased that performance by almost 10 percentage points. Furthermore, we observed that the model trained with 100 % of the training partition reached a state-of-the-art level of performance with an overall accuracy of 0.869 ± 0.064, 95 % CI: $[0.865 - 0.872]$ and Cohen's κ of 0.799 ± 0.098, 95 % CI: $[0.794 - 0.804]$ (Table 3.9). The model furthermore performs well with respect to classifying individual sleep stages as shown in the confusion matrix in Figure 3.6b. However, the model still has difficulties classifying and distinguishing between certain sleep stages, especially between N2, N1, and N3; and W, N2, and N1.

3.3.3.5 *Increasing the number of data sources improves classification performance*

On average, we saw an increase in overall accuracy, when increasing the number of cohorts from 2 to 4 using 500 PSGs in each configuration, see Fig-

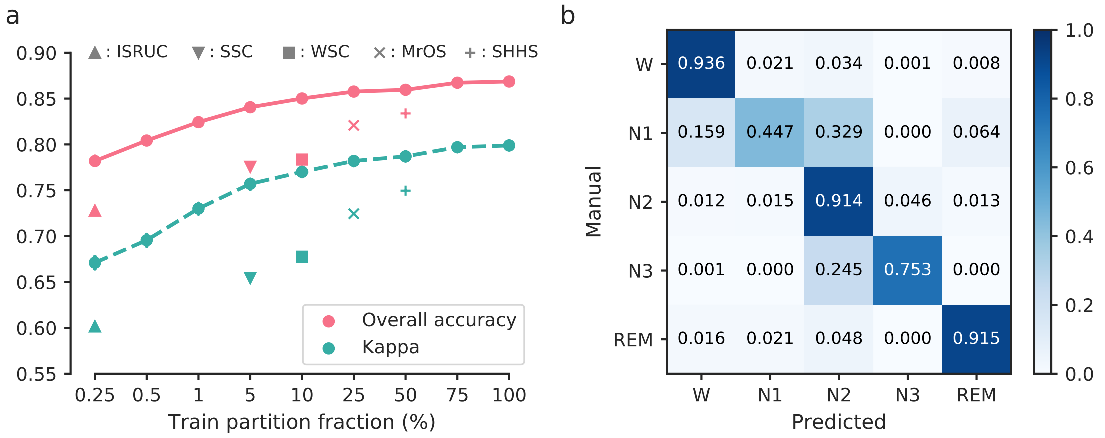


Figure 3.6: Training on mixed data increased predictive performance compared to individual cohorts of similar size. a) There is a gain in predictive performance by mixing data from various sources consistent across the size of the training dataset. b) Confusion matrix for a model trained on 100% of the available training partition data. The model shows excellent performance overall, with most misclassification happening between W and N1, and N1, N2, and N3. This is somewhat consistent with clinical experience, since N1 is a transition stage between wake and the deeper stages of sleep with much frequency content overlap with both W and N2.

ure 3.7 and Table 3.10. Specifically, we found that the average overall accuracy increased from 0.788 ± 0.102, 95 % CI: [0.787 − 0.790] in the 2-cohort configuration to 0.808 ± 0.092, 95 % CI: [0.807 − 0.810] and 0.821 ± 0.085, 95 % CI: [0.819 − 0.823] in the 3- and 4-cohort configurations, respectively.

3.3.4 *Discussion*

In this work, we present an end-to-end deep learning-based model for fully automatic micro- and macro-sleep stage classification. Using all of the available data sources for training our model, we reached an overall accuracy on test partition of 0.869 ± 0.064, 95 % CI: [0.865 − 0.872], and a Cohen's κ of 0.799 ± 0.098, 95 % CI: [0.794 − 0.804], which is in the very high end of the substantial agreement category for observer agreement [120]. We found that individual cohorts exhibit major differences in overall accuracy and Cohen's κ when subjected to both training and testing conditions and specifically, we found that average performance on the test partition in the LOCI configurations varied significantly from 0.676 ± 0.124, 95 % CI: [0.670 − 0.682] when training on ISRUC, to 0.837 ± 0.084, 95 % CI: [0.833 − 0.841] when training on SHHS. Each individual cohort also showed large deviations in predictive performance when tested on the other cohorts. For example, when conditioned on SHHS data, the lowest average accuracy was 0.721 on SSC test data compared to the highest at 0.872 on SHHS test data, while conditioning on SSC training data, the lowest average accuracy was 0.704 on ISRUC test data compared to 0.824 on WSC test data. Classification performance was generally higher on the test set when using the LOCO configuration, except for SHHS (higher in LOCI) and SSC (no difference). We also found that having data from multiple sources always resulted in better-performing models compared to training on single cohorts. Increasing the number of data sources increased classification performance, although this was non-significant. In the design of the model, we observed that model performance was enhanced by the addition of the recurrent module (bGRU), a phenomenon likely reflect-

Table 3.8: Performance characteristics for LOCI and LOCO training configurations.

		Overall accuracy				Cohen's kappa			
	N PSGs	**Mean**	**SD**	**Median**	**95% CI, mean**	**Mean**	**SD**	**Median**	**95% CI, mean**
LOCI-wd									
ISRUC	1584	0.679	0.123	0.701	[0.673-0.685]	0.542	0.169	0.574	[0.533-0.550]
MrOS	1584	0.821	0.077	0.835	[0.817-0.825]	0.727	0.114	0.745	[0.721-0.733]
SHHS	1584	0.834	0.088	0.858	[0.830-0.839]	0.750	0.132	0.786	[0.744-0.757]
SHHS	1584	0.762	0.094	0.774	[0.757-0.767]	0.639	0.129	0.654	[0.633-0.646]
WSC	1584	0.758	0.105	0.773	[0.753-0.764]	0.633	0.145	0.653	[0.626-0.640]
LOCI									
ISRUC	1584	0.676	0.124	0.700	[0.670-0.682]	0.539	0.170	0.574	[0.531-0.547]
MrOS	1584	0.826	0.074	0.839	[0.822-0.829]	0.732	0.111	0.748	[0.726-0.737]
SHHS‡	1584	0.837	0.084	0.858	[0.833-0.841]	0.754	0.127	0.786	[0.748-0.761]
SHHS	1584	0.773	0.088	0.785	[0.769-0.777]	0.657	0.125	0.671	[0.651-0.663]
WSC	1584	0.763	0.101	0.776	[0.758-0.768]	0.641	0.140	0.659	[0.635-0.648]
LOCO									
ISRUC†	52	0.749	0.081	0.764	[0.727-0.771]	0.648	0.119	0.682	[0.616-0.680]
	126	0.757	0.071	0.766	[0.744-0.769]	0.661	0.101	0.682	[0.643-0.678]
MrOS†	371	0.843	0.066	0.851	[0.836-0.849]	0.757	0.104	0.776	[0.746-0.767]
	3932	0.841	0.069	0.854	[0.838-0.843]	0.752	0.107	0.775	[0.749-0.755]
SHHS	846	0.805	0.076	0.815	[0.800-0.810]	0.705	0.109	0.722	[0.698-0.712]
	8444	0.800	0.081	0.811	[0.798-0.801]	0.697	0.115	0.713	[0.694-0.699]
SHHS	76	0.793	0.086	0.809	[0.744-0.812]	0.680	0.120	0.700	[0.653-0.707]
	766	0.798	0.086	0.815	[0.792-0.805]	0.690	0.123	0.711	[0.681-0.699]
WSC†	239	0.826	0.065	0.835	[0.818-0.834]	0.720	0.096	0.736	[0.708-0.732]
	2411	0.824	0.068	0.837	[0.821-0.827]	0.718	0.100	0.736	[0.714-0.722]

Metrics are aggregated across all subjects for each cohort in test partition (N = 1584 PSGs). Bottom rows in LOCO configuration correspond to evaluating performance on entire cohort. PSG: polysomnography; LOCI: leave-one-cohort-in; LOCI-wd: LOCI with weight decay; LOCO: leave-one-cohort-out; ISRUC: Institute of Systems and Robotics, University of Coimbra Sleep Cohort; MrOS: Osteoporotic Fractures in Men Sleep Study; SHHS: Sleep Heart Health Study; SSC: Stanford Sleep Cohort; WSC: Wisconsin Sleep Cohort; †significantly better than corresponding LOCI; ‡significantly better than corresponding LOCO.

ing the fact that sleep stage scoring at a specific time in one subject can be influenced by signal content (frequency, amplitude, presence of micro-events) at later time steps. However, the complexity of the module given by the number of hidden units did not affect performance. In all our experiments, we also evaluated the performance of the model every 1 s compared to the performance evaluated every 30 s and found them to be similar, which indicates the model is stable in classification in periods corresponding to an epoch of data.

Only a handful of studies have previously reported results when using multiple cohorts [73], [89], [90]. Some authors have reported a drop from 81.9% to 77.7% when training on the Massachusetts General Hospital cohort (MGH) and testing on MGH and SHHS, respectively [89], while others have shown significant drops from 89.8% to 81.4% and 72.1% on two separate hold-out sets from Singapore and USA [90]. We also observed similar trends in our LOCI and LOCO experiments, where excluding the training subset of a cohort from the training partition resulted in a significant drop in performance on the respective test subset from that cohort. A benefit of

Table 3.9: Model performance of test partition with varying fractions of training data.

	Overall accuracy				Cohen's kappa			
Fraction, %	**Mean**	**SD**	**Median**	**95% CI, mean**	**Mean**	**SD**	**Median**	**95% CI, mean**
0.25	0.782	0.097	0.801	[0.777-0.787]	0.671	0.141	0.696	[0.664-0.678]
0.50	0.804	0.086	0.824	[0.800-0.808]	0.696	0.131	0.724	[0.689-0.702]
1	0.824	0.079	0.840	[0.820-0.828]	0.730	0.118	0.753	[0.724-0.736]
5	0.841	0.074	0.856	[0.837-0.844]	0.757	0.113	0.780	[0.751-0.763]
10	0.850	0.069	0.864	[0.847-0.853]	0.770	0.108	0.791	[0.765-0.775]
25	0.858	0.066	0.873	[0.854-0.861]	0.782	0.102	0.804	[0.777-0.787]
50	0.860	0.063	0.874	[0.856-0.863]	0.787	0.097	0.809	[0.782-0.792]
75	0.867	0.062	0.882	[0.864-0.870]	0.797	0.096	0.818	[0.792-0.802]
100	0.869	0.064	0.883	[0.865-0.872]	0.799	0.098	0.820	[0.794-0.804]

Increasing the available training data increased performance on the test partition ($N = 1584$) shown here as aggregated metrics across all subjects. No statistical difference was found by comparing confidence intervals between models trained with 75% and 100% of available training data, which indicates a saturation in training.

our LOCI and LOCO experiments is the possibility for direct benchmarking against previous publications using specific cohorts in their experiments. For example, we obtain an accuracy of 0.805 in the LOCO-SHHS training-testing case compared to 0.777 previously reported by Biswal *et al.* [89], both of which reflect classification performance when SHHS had not been used for training; and an accuracy of 0.865 in the LOCI-WSC case compared to 0.841 reported previously [72], where both have been using a subset of WSC for training the model. Interestingly, we obtained the same level of performance on the SHHS data in our LOCI experiment as reported by Sors *et al.* (87% accuracy, 81% Cohen's κ) even though they only used single-EEG for their experiments [121]. Other works that have investigated single- vs. multi-channel models for automatic sleep stage classification have found that models generally benefit from having more channels available for training [85], [87], [89]. It may be that some cohorts share different characteristics that makes them more suitable for single- or multi-channel models, but this is speculative and would need to be verified in subsequent studies.

We only optimized our network architecture with respect to the temporal processing module and therefore cannot assess what impact different design choices for the other modules would have had on final performance. For example, the EMG signal has different statistical properties and spectral content, and separate, parallel architectures for EMG and EEG/ EOG feature extraction may be warranted, as proposed by others [73], [85]. Other studies have however shown equal performance in large cohorts using a similar channel mixing approach as proposed here [72]. Another limitation is found in our training runs, as we did not consider balancing our data with respect to the proportion of sleep stages, which may or may not have had impact on overall performance. It is well established that there is significant variation in scoring and validation of N1/REM and N2/N3 [24], [28], [29], which challenges the training for any classification algorithm. Some researchers have experimented balancing the cost of misclassifying sleep stages by weighting them by their inverse frequency of occurrence and found

Table 3.10: Model performance on test partition (N = 1584) with varying number of cohorts in training partition.

	Overall accuracy				Kappa			
Training cohorts	**Mean**	**SD**	**Median**	**95% CI, mean**	**Mean**	**SD**	**Median**	**95% CI, mean**
2								
Overall	0.788	0.102	0.811	[0.787-0.790]	0.683	0.143	0.710	[0.681-0.685]
ISRUC-MrOS	0.781	0.102	0.804	[0.776-0.786]	0.675	0.143	0.703	[0.668-0.682]
ISRUC-SHHS	0.808	0.097	0.835	[0.804-0.813]	0.717	0.142	0.756	[0.710-0.724]
ISRUC-SSC	0.735	0.103	0.753	[0.729-0.740]	0.613	0.140	0.638	[0.606-0.620]
ISRUC-WSC	0.745	0.107	0.758	[0.740-0.750]	0.628	0.140	0.642	[0.621-0.635]
MrOS-SHHS	0.829	0.081	0.849	[0.825-0.833]	0.740	0.124	0.769	[0.734-0.746]
MrOS-SSC	0.796	0.090	0.816	[0.791-0.800]	0.683	0.133	0.708	[0.677-0.690]
MrOS-WSC	0.805	0.087	0.822	[0.801-0.809]	0.699	0.126	0.722	[0.693-0.705]
SHHS-SSC	0.816	0.090	0.839	[0.812-0.821]	0.722	0.129	0.755	[0.716-0.729]
SHHS-WSC	0.824	0.089	0.846	[0.820-0.828]	0.733	0.128	0.762	[0.727-0.739]
SSC-WSC	0.742	0.110	0.755	[0.737-0.748]	0.620	0.145	0.634	[0.613-0.627]
3								
Overall	0.808	0.092	0.830	[0.807-0.810]	0.711	0.131	0.739	[0.709-0.713]
ISRUC-MrOS-SHHS	0.820	0.092	0.844	[0.815-0.825]	0.732	0.134	0.766	[0.725-0.738]
ISRUC-MrOS-SSC	0.798	0.088	0.816	[0.794-0.802]	0.694	0.129	0.720	[0.688-0.700]
ISRUC-MrOS-WSC	0.811	0.083	0.828	[0.807-0.815]	0.711	0.119	0.735	[0.705-0.717]
ISRUC-SHHS-SSC	0.807	0.090	0.828	[0.803-0.812]	0.714	0.126	0.739	[0.708-0.721]
ISRUC-SHHS-WSC	0.817	0.091	0.842	[0.813-0.822]	0.728	0.128	0.759	[0.722-0.735]
ISRUC-SSC-WSC	0.755	0.109	0.775	[0.750-0.760]	0.639	0.150	0.670	[0.631-0.646]
MrOS-SHHS-SSC	0.833	0.071	0.848	[0.829-0.837]	0.744	0.109	0.766	[0.739-0.750]
MrOS-SHHS-WSC	0.840	0.073	0.854	[0.836-0.843]	0.753	0.109	0.774	[0.748-0.759]
MrOS-SSC-WSC	0.795	0.088	0.811	[0.791-0.800]	0.687	0.123	0.706	[0.681-0.693]
SHHS-SSC-WSC	0.807	0.101	0.833	[0.802-0.812]	0.710	0.142	0.744	[0.703-0.717]
4								
Overall	0.821	0.085	0.840	[0.819-0.823]	0.728	0.124	0.755	[0.726-0.731]
ISRUC-MrOS-SHHS-SSC	0.827	0.078	0.843	[0.823-0.831]	0.739	0.115	0.764	[0.733-0.744]
ISRUC-MrOS-SHHS-WSC	0.835	0.075	0.850	[0.831-0.838]	0.747	0.112	0.768	[0.742-0.753]
ISRUC-MrOS-SSC-WSC	0.794	0.097	0.817	[0.789-0.799]	0.687	0.139	0.716	[0.680-0.694]
ISRUC-SHHS-SSC-WSC	0.819	0.091	0.843	[0.814-0.823]	0.728	0.131	0.759	[0.721-0.734]
MrOS-SHHS-SSC-WSC	0.830	0.076	0.846	[0.826-0.834]	0.741	0.112	0.763	[0.736-0.747]

The total number of training records were fixed at N = 500 for all configurations. ISRUC: Institute of Systems and Robotics, University of Coimbra Sleep Cohort; MrOS: Osteoporotic Fractures in Men Sleep Study; SHHS: Sleep Heart Health Study; SSC: Stanford Sleep Cohort; WSC: Wisconsin Sleep Cohort.

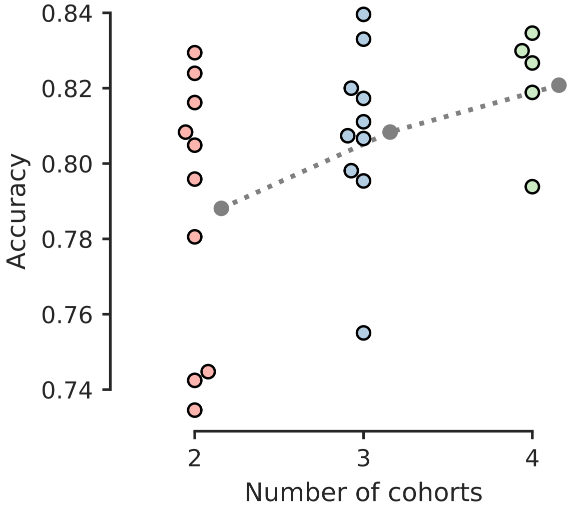


Figure 3.7: Number of cohorts in training partition increases model performance. Each datapoint is shown as the overall accuracy aggregated across all subjects for a specific training configuration. For example, the bottom dot in column 2 (3 cohort configuration) shows the performance on the test set (overall accuracy 0.755 ± 0.109, 95 % CI: [0.750 – 0.760]), when training with 500 PSGs randomly and evenly drawn from SSC, ISRUC, and WSC. Notice the scale on the y-axis.

no significant improvement [72], [121], while others have experimented with balancing the sleep stage frequencies in each batch of data input to the neural network model [85], but more rigorous research in resampling or over/under-sampling techniques is warranted in this regard. We ultimately decided against experimenting with balancing our sleep stages in each batch, as we prioritized flexibility with regards to the length of input sequences fed to the network. All our models ran through at least 50 epochs of training (passes through the training partition), which might have induced a bias in the configurations with larger cohorts. For example, one pass through the training partition in the LOCI-ISRUC case corresponds to much less data than one pass through the LOCI-SHHS case. However, since we selected the best performing model based on Cohen's κ across all 50 epochs, we have allowed for more effective training in cases with less available training data. We observed that models using less data in the training partition generally had to run for longer time (i. e. more epochs) before converging.

In future studies on automatic sleep stage classification algorithms, we strongly recommend researchers to test and report results on not just hold-out test partitions, but also on cohorts completely unseen by the model both during training and testing/validation. Our experiments indicate that even though good performance can be achieved on hold-out data using a single cohort, this does not necessarily translate into good generalization performance. Such approach requires availability of many publicly available, high-quality, well-documented databases with easily accessible PSG data, associated annotations and related patient information. In this regard, websites such as the NSRR, which contains several large databases with clinical data as well as PSG and annotation data in a standardized format [107], [108], are an invaluable resource for researchers. We also propose that the sleep science community establishes a common reference dataset on which researchers in machine learning can benchmark their models, similar to what the computer vision and general machine learning community has done with the ImageNet Large Scale Visual Recognition Challenge [122], an annual competition in which researchers submit their models to test in various competitions.

In summary, we have developed an automatic sleep stage classification algorithm based on deep learning, that can accurately classify sleep stages at a flexible resolution with a state-of-the-art classification performance of 87% accuracy on a test set of 1584 PSGs. We trained and tested our model using five cohorts with varying numbers of PSGs covering multiple phenotypes with specific focus on how well cohorts can generalize to each other. We found that different cohorts generalize very differently both in intra- and inter-cohort settings (LOCI vs. LOCO experiments). Furthermore, we also found that having more data sources significantly improve classification performance and generalizability to the extent that even just a small number of training PSGs can reach high classification performance by including many different sources. To our knowledge, this is one of the largest, if not the largest, study on automatic sleep stage classification in terms of PSG volume, diversity, and performance.

3.4 PAPER III: NEURAL NETWORK ANALYSIS OF SLEEP STAGES ENABLES EFFICIENT DIAGNOSIS OF NARCOLEPSY

ABSTRACT: Analysis of sleep for the diagnosis of sleep disorders such as NT1 currently requires visual inspection of polysomnography records by trained scoring technicians. Here, we used neural networks in approximately 3000 normal and abnormal sleep recordings to automate sleep stage scoring, producing a hypnodensity graph—a probability distribution conveying more information than classical hypnograms. Accuracy of sleep stage scoring was validated in 70 subjects assessed by six scorers. The best model performed better than any individual scorer (87% versus consensus). It also reliably scores sleep down to 5 s instead of 30 s scoring epochs.

3.4.1 *Materials & Methods*

3.4.1.1 *Datasets*

The success of machine learning depends on the size and quality of the data on which the model is trained and evaluated [123], [124]. We used a large dataset comprised of several thousand sleep studies to train, validate, and test/replicate our models. To ensure heterogeneity, data came from 4 different cohorts: SSC [111], [112], WSC [112], [125], Interscorer Reliability Cohort (IS-RC) [126], and Korean Hypersomnia Cohort (KHC) [127] Institutional Review Boards approved the study and informed consent was obtained from all participants. Technicians trained in sleep scoring manually labeled all sleep studies. Figure 3.8a and b summarize the overall design of the study for sleep stage scoring. Table 3.11 provides a summary of the size of each cohort and how it was used. For this analysis, a few recordings with poor quality sleep studies, i.e. missing critical channels, with additional sensors or with a too short sleep duration ($\leqslant$ 2 h) were excluded. Below is a brief description of each dataset.

POPULATION-BASED WISCONSIN SLEEP COHORT This cohort is a longitudinal study of state agency employees aged 37–82 years from Wisconsin, USA, and approximates a population-based sample (see Table 3.11 for age at study). The subjects in this study are generally more overweight [125]. The study is ongoing, and dates back to 1988. 2167 PSGs in 1086 subjects were used for training while 286 randomly selected PSGs were used for validation testing of the sleep stage-scoring algorithm.Approximately 25% of the population have an AHI above 15 h and 40% have a periodic leg movement index (PLMI) above 15 h. A detailed description of the sample can be found in [125] and [112].

PATIENT-BASED STANFORD SLEEP COHORT PSGs from this cohort were recorded at the Stanford Sleep Clinic dating back to 1999, and represent sleep disorder patients aged 18-91 visiting the clinic (see Table 3.11 for age at study). The cohort contains thousands of PSG recordings, but for this study we used 894 diagnostic (no positive airway pressure (PAP)) recordings in independent patients that have been used in prior studies [126]. This subset contains patients with a range of different diagnoses including: sleep

Table 3.11: Description of the various cohorts included in this study.

Cohort	Age, years	BMI, kg m^{-2}	Sex, %	Train	Test
WSC	59.7 ± 8.4	31.6 ± 7.1	53.1	1086 (2167)	286
SSC	45.4 ± 13.8	23.9 ± 6.5	59.4	617	277
KHC	29.1 ± 13.2	24.1 ± 4.3	58.6	None	160
IS-RC	51.1 ± 4.2	32.9 ± 9.2	0	None	70
Total subjects				1703	793
Total PSGs				2784	793

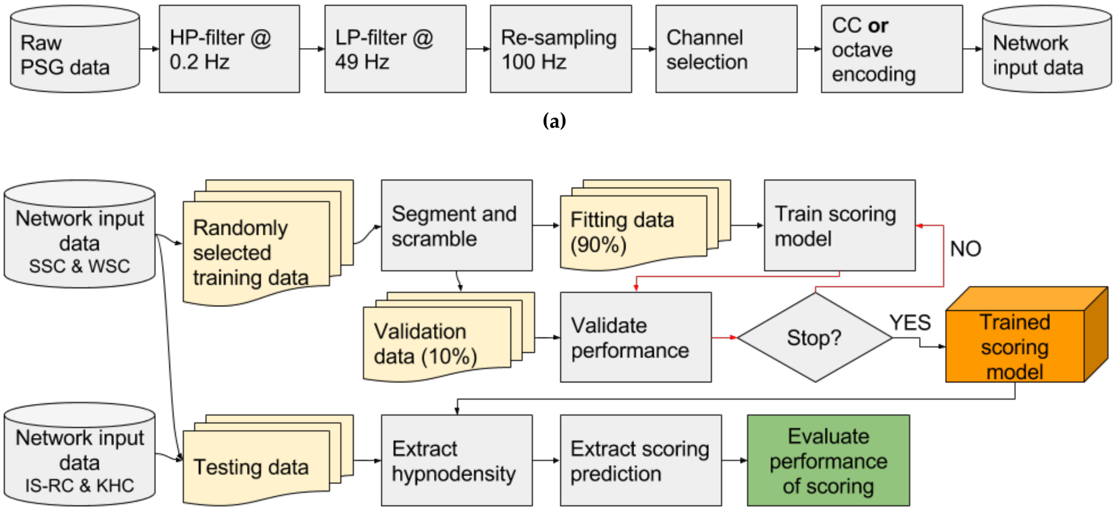
Variables are aggregated across PSGs. WSC: Wisconsin Sleep Cohort; SSC: Stanford Sleep Cohort; KHC: Korean Hypersomnia Cohort; IS-RC: Interscorer Reliability Cohort.

disordered breathing (607), insomnia (141), REM sleep behavior disorder (4), restless legs syndrome (23), NT1 (25), delayed sleep phase syndrome (14), and other conditions (39). Description of the subsample can be found in [111] and [112]. Approximately 30% of subjects have an AHI above 15 h, or a PLMI above 15 h. 617 randomly selected subjects were used for training the neural networks while 277 randomly selected PSGs were kept for validation testing of the sleep stage scoring algorithm. 26 subjects were removed from the study—4 due to poor data quality, and the rest due to continued medication usage.

patient-based korean hypersomnia cohort The Korean Hypersomnia Cohort is a high pretest probability sample for narcolepsy. It includes 160 patients with a primary complaint of excessive daytime sleepiness (see Table 3.11 for age at study). These PSGs were used for testing the sleep scoring algorithm.No data was used for training the sleep-scoring algorithm. Detailed description of the sample can be found in [127] and [111].

patient-based inter-scorer reliability cohort As Rosenberg and Van Hout [26] have shown, variation between individual scorers can sometimes be large, leading to an imprecise gold standard. To quantify this, and to establish a more accurate gold standard, 10 scorers from five different institutions, University of Pennsylvania, St. Luke's Hospital, University of Wisconsin at Madison, Harvard University, and Stanford University, analyzed the same 70 full night PSGs. This allowed for a much more precise gold standard, and the inter-scorer reliability could be quantified for a dataset, which could also be examined by automatic scoring algorithms. For this study, scoring data from University of Pennsylvania, St. Luke's and Stanford were used. All subjects are female (see Table 3.11 for details). Detailed description of the sample can be found in [126] and [128].

american academy of sleep medicine sleep study The AASM ISR dataset is composed of a single control sleep study of 150 epochs of each 30 s in length that was scored by 5234 ± 14 experienced sleep technologists for quality control purposes. Design of this dataset is described in [26].



(a)

(b)

Figure 3.8: Overview of STAGES model for sleep stage classification. (a) Pre-processing steps taken to achieve the format of data as it is used in the neural networks. One of the 5 channels is first high-pass filtered with a cut-off at 0.2 Hz, then low-pass filtered with a cut-off at 49 Hz followed by a re-sampling to 100 Hz to ensure data homogeneity. In the case of EEG signals, a channel selection is employed to choose the channel with the least noise. The data are then encoded using either the cross-correlation (CC) or the octave encoding. (b) Producing and testing the automatic scoring algorithm. A part of the SSC [111], [112] and WSC [112], [125] is randomly selected, as described in Table 3.11. These data are then segmented in 5 min segments and scrambled with segments from other subjects to increase batch similarity during training. A neural network is then trained until convergence (evaluated using a separate validation sample). Once trained, the networks are tested on a separate part of the SSC and WSC along with data from the IS-RC [126] and KHC [111], [127].

3.4.1.2 *Data labels, scoring and fuzzy logic*

Sleep stages were scored by PSG-trained technicians using established scoring rules, as described in the AASM Scoring Manual [129]. In doing so, technicians assign each epoch with a discrete value. With a probabilistic model, like the one proposed in this study, a relationship to one of the fuzzy sets is inferred based on thousands of training examples labeled by many different scoring-technicians.

The hypnodensity graph refers to the probability distribution over each possible stage for each epoch, as seen in Figures 3.14 and 3.15. This allows more information to be conveyed, since every epoch of sleep within the same stage is not identical. For comparison with the gold standard, however, a discrete value must be assigned from the model output as:

$$\hat{y} = \arg\max_{y_i} \sum_{i}^{N} \mathbf{P}_i(\mathbf{y}_i \,|\, \mathbf{x}_i), \tag{3.20}$$

where $\mathbf{P}_i(\mathbf{y}_i \,|\, \mathbf{x}_i)$ is a vector with the estimated probabilities for each sleep stage in the ith segment, N is the number of segments an epoch is divided into, and $\hat{y}$ is the estimated label.

Sleep scoring technicians score sleep in 30 s epochs, based on what stage they assess is represented in the majority of the epoch—a relic of when recordings were done on paper. This means that when multiple sleep stages are represented, more than half of the epoch may not match the assigned label. This is evident in the fact that the label accuracy decreases near transition epochs [26]. One solution to this problem is to remove transitional regions to purify each class. However, this has the disadvantage of under-sampling transitional stages, such as N1, and removes the context of quickly changing stages, as is found in a spontaneous arousal. It has been demonstrated that the negative effects of imperfect "noisy" labels may be mitigated if a large enough training dataset is incorporated and the model is robust to overfitting [130]. This also assumes that the noise is randomly distributed with an accurate mean—a bias cannot be canceled out, regardless of the amount of training data. For these reasons, all data including those containing sleep transitions were included. Biases were evaluated by incorporating data from several different scoring experts cohorts and types of subjects.

To ensure quick convergence, while also allowing for long-term dependencies in memory-based models, the data were broken up in 5 min blocks and shuffled to minimize the shift in covariates during training caused by differences between subjects. To quantify the importance of segment sizes, both 5 s and 15 s windows were also tested.

3.4.1.3 *Data selection and pre-processing*

A full night PSG PSG involves recording many different channels, some of which are not necessary for sleep scoring [131]. In this study, EEG (C3 or C4, and O1 or O2), chin EMG, and the left and right EOG channels were used, with reference to the contralateral mastoid. Poor electrode connections are common when performing a PSG analysis. This can lead to a noisy recording, rendering it useless. To determine whether right or left EEG channels were used, the noise of each was quantified by dividing the EEG data in 5 min segments, and extracting the Hjorth parameters [132]. These were then log-transformed, averaged, and compared with a previously established multivariate distribution, based on the WSC [112], [125] and SSC [111], [112] training data. The channel with lowest Mahalanobis distance to this distribution was selected. The log-transformation has the advantage of making flat signals/disconnects as uncommon as very noisy signals, in turn making them less likely to be selected. To minimize heterogeneity across recordings, and at the same time reducing the size of the data, all channels were down-sampled to 100 Hz. Additionally, all channels were filtered with a 5th order two-directional infinite impulse response (IIR) high-pass filter with cutoff frequency of 0.2 Hz and a 5th order two-directional IIR low-pass filter with cutoff frequency of 49 Hz. The EMG signal contains frequencies well above 49 Hz, but since much data had been down-sampled to 100 Hz in the WSC, this cutoff was selected for all cohorts. All steps of the pre-processing are illustrated in Figure 3.8a.

The AASM recommends EEG, EOG, and chin EMG for sleep stage scoring, see Section 2.2.1.

3.4.1.4 *Convolutional and recurrent neural networks*

convolutional neural networks (CNNs) are a class of deep learning models initially developed to solve problems in the field of computer vision [33]. A CNN is a machine learning model in which a high-dimensional input, such as an image, is transformed through a network of filters and sub-sampling layers. Each layer of filters produces a set of features from the previous layer, and as more layers are stacked, more complex features are generated. This network is coupled with a general-purpose learning algorithm, resulting in features produced by the model reflecting latent properties of the data rather than the imagination of the designer. This property places fewer constrictions on the model by allowing more flexibility, and hence the predictive power of the model will increase as more data is observed. This is facilitated by the large number of parameters in such a model, but may also necessitate a large amount of training data.

Sleep stage scoring involves a classification of a discrete time-series, in which adjacent segments are correlated. Models that incorporate memory may take advantage of this and may lead to better overall performance by evening out fluctuations. However, these fluctuations may be the defining trait or anomaly of some underlying pathology present in only a fraction of subjects, and perhaps absent in the training data. This can be thought of similarly to a person with a speech impediment: the contextual information will ease the understanding, but knowing only the output, this might also hide the fact that the person has such a speech impediment. To highlight the importance of this fact, models with and without memory were applied in this work. Memory can be added to such a model by introducing recurrent connections in the final layers of the model. This turns the model into a recurrent neural network (RNN). Classical RNNs had the problem of vanishing or exploding gradients, which meant that optimization was very difficult. This problem was solved by changing the configuration of the simple hidden node into an long short-term memory (LSTM) cell [133]. Models without this memory are referred to as FF models. A more in-depth explanation of CNNs including application areas can be found the review article on deep learning by LeCun, Bengio, and Hinton [33], and the deep learning textbook by Goodfellow, Bengio, and Courville [134]. For a more general introduction to machine learning concepts, see the textbook on pattern recognition and machine learning by Bishop [135].

Narcolepsy is one such pathology well known to involve abnormal sleep stages transitions.

These models have no recurrency and thus feed forward the signals directly, hence FF.

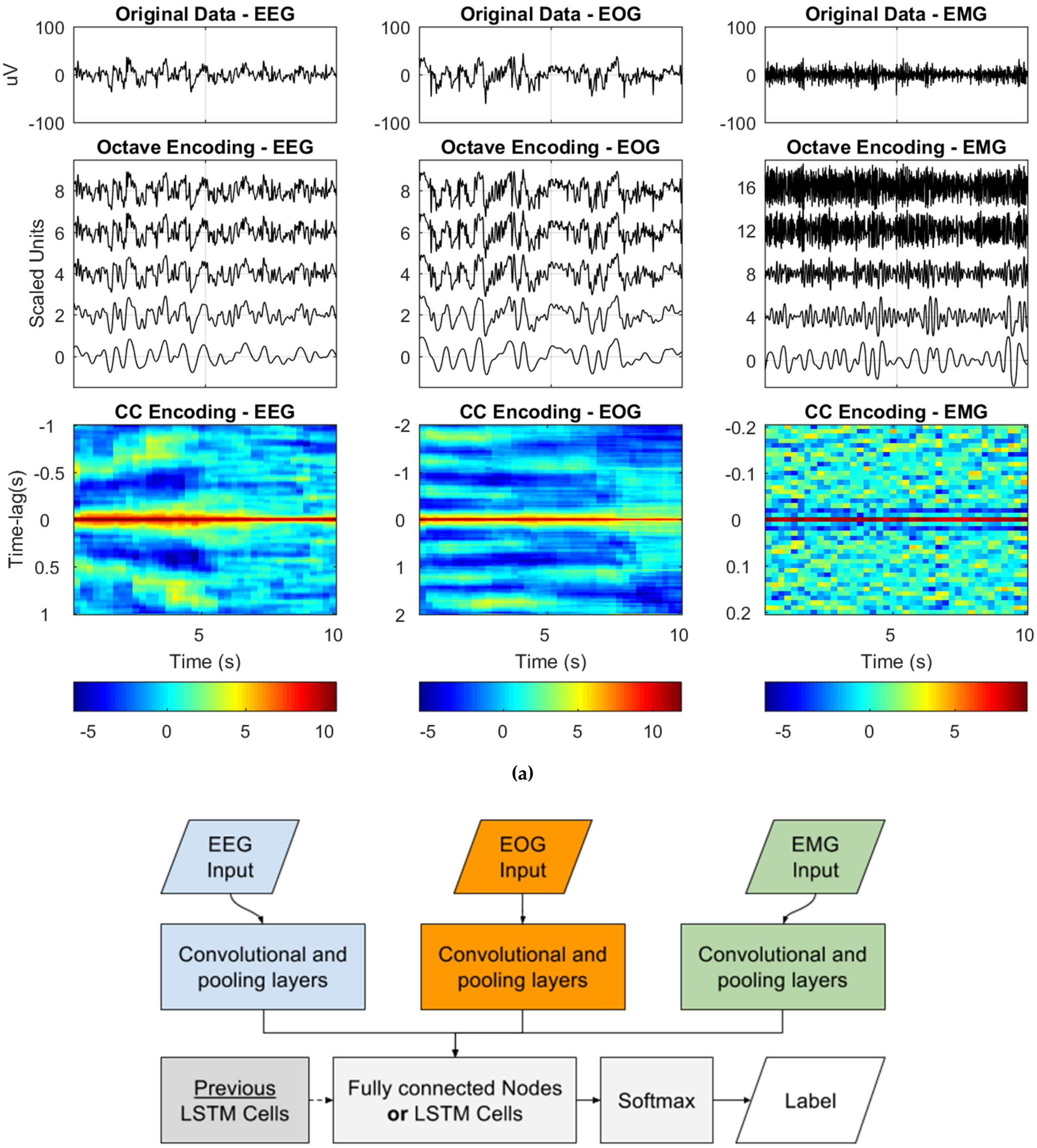
3.4.1.5 *Data input and transformations*

Biophysical signals, such as those found in a PSG, inherently have a low signal to noise ratio, the degree of which varies between subjects, and hence learning robust features from these signals may be difficult. To circumvent this, two representations of the data that could minimize these effects were selected. An example of each decomposition is shown in Figure 3.9.

Octave encoding maintains all information in the signal, and enriches it by repeatedly removing the top half of the bandwidth (i.e. cut off frequencies of 49 Hz, 25 Hz, 12.5 Hz, 6.25 Hz and 3.125 Hz) using a series of low-pass filters, yielding a total of 5 new channels for each original channel. At no point is a high-pass filter applied. Instead, the high frequency information may be obtained by subtracting lower frequency channels—an association the neural networks can make, given their universal approximator properties [136]. After filtration, each new channel is scaled to the 95th percentile and log modulus transformed:

$$\mathbf{x}_{\text{scaled}} = \text{sign}(\mathbf{x}) \log\left(\frac{|\mathbf{x}|}{p_{95}(\mathbf{x})} + 1\right) \tag{3.21}$$

The initial scaling places 95th of the data between -1 and 1, a range in which the log modulus is close to linear. Very large values, such as those found in particularly noisy areas, are attenuated greatly. Some recordings are noisy, making the 95th percentile significantly higher than what the physiology reflects. Therefore, instead of the selecting the 95th percentile from the entire recording, the recording is separated into 50% overlapping 90 min segments, from which the 95thth percentile is computed. The mode of these values is then used as a scaling reference. In general, scaling and normalization is important to ensure quick convergence as well as generalization in neural



(a)

(b)

Figure 3.9: Neural network strategy for STAGES model. (a) An example of the octave and the CC encoding on 10 s of EEG, EOG and EMG data. These processed data are fed into the neural networks in one of the two formats. The data in the octave encoding are offset for visualization purposes. Color scale is unitless. (b) Simplified network configuration, displaying how data are fed and processed through the networks. A more detailed description of the network architecture is shown in Figure 3.11.

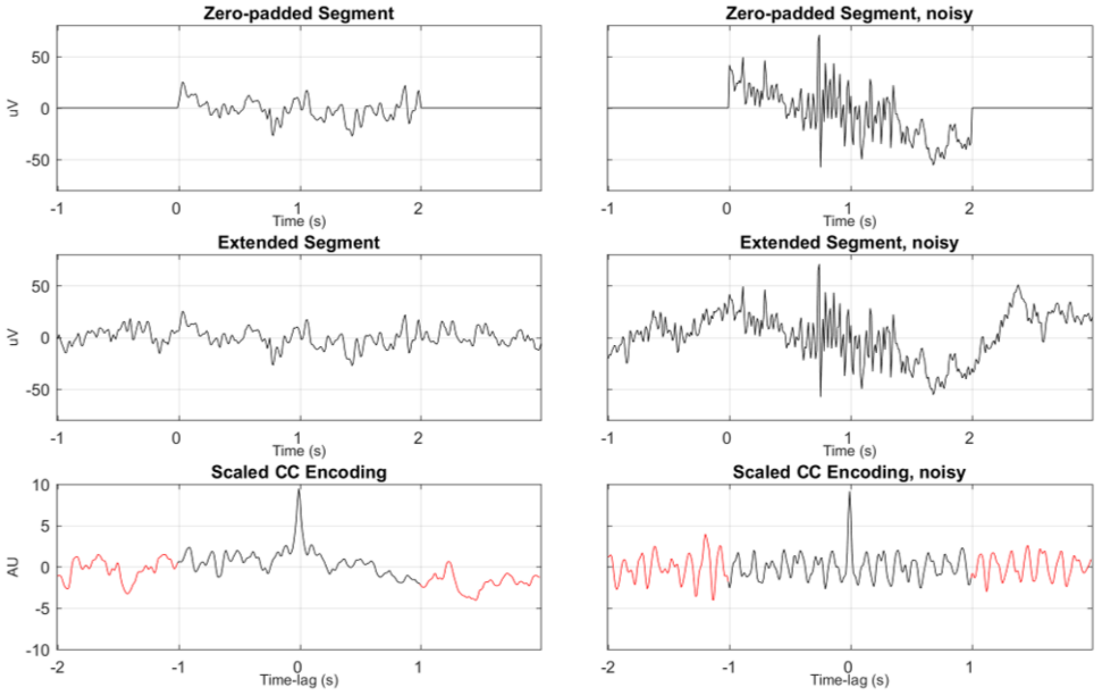


Figure 3.10: Implementation of CC encoding. CC encoding of a noisy (right) and less noisy (left) signal. The central part of the encoding, representing areas of full overlap between correlated signals, is kept; the red part is discarded.

networks. The decomposition is done in the same way on every channel, resulting in 25 new channels in total.

Using a CC function, underlying periodicities in the data are revealed while noise is attenuated. White noise is by definition uncorrelated; the auto-correlation function is zero everywhere except in lag zero. It is this property that is utilized, even though noise cannot always be modeled as such. PSG signals are often obscured by undesired noise that is uncorrelated with other aspects of the signals. An example CC between a signal segment and an augmented version of the same signal segment is shown in Figure 3.10.

Choosing the CC in this manner over a standard auto-correlation function serves two purposes: the slow frequencies are expressed better, since there is always full overlap between the two signals; and the change in fluctuations over time within a segment is expressed, making the function reflect aspects of stationarity. Because this is the CC between a signal and an augmented version of itself, the zero lag represents the power of that segment, as is the case in an auto-correlation function.

Although some of this can be adjusted with the normal auto-correlation function using an unbiased estimate.

Frequency content with a time resolution may also be expressed using time-frequency decompositions, such as spectrograms or scalograms. One of the key properties of a CNN, however, is the ability to detect distinct features anywhere in an input given the latent property of equivariance [137]. A CC function reveals an underlying set of frequencies as an oscillation pattern, as opposed to a spectrogram, where frequencies are displayed as small streaks or spots in specific locations, corresponding to time-specific frequencies. The length and size of each CC reflects the expected frequency content and the limit of quasi-stationarity.

That is, how quickly the frequency content is expected to change.

The EEG signal is quasi-stationary in signals with a length of up to 0.25 s [138], [139]. The lowest expected meaningful frequencies are delta rhythms, which have a lower bound of 0.5 Hz [138]. Hence, the transformation is made up of 2 s segments with 1.75 s overlaps between segments.

The EOG signal reveals information about eye movements such as REMs, and to some extent EEG activity [128], [129]. In the case of the EOG signal, the relative phase between the two channels is of great importance to determine synchronized eye movements, and hence a CC of opposite channels is also included. The slowest eye-movements happen over the course of several seconds [128], [129], and hence a segment length of 4 s was selected for the correlation functions. To maintain resolution flexibility with the EEG, an overlap of 3.75 s between each data segment was selected.

Either the augmented or zero-padded signal is replaced with the opposite channel.

In the case of the EMG signal, the main concern is the signal amplitude and the temporal resolution, not the actual frequencies. As no relevant low frequency content is expected, a segment length of 0.4 seconds and an overlap of 0.25 seconds was selected.

As with the octave encoding, the data is scaled, although only within segments:

$$D_i = \frac{\gamma_{\mathbf{x}_i\mathbf{y}_i} \log(1 + \max|\gamma_{\mathbf{x}_i\mathbf{y}_i}|)}{\max|\gamma_{\mathbf{x}_i\mathbf{y}_i}|} \tag{3.22}$$

where D_i is the scaled correlation function and $\gamma_{\mathbf{x}_i\mathbf{y}_i}$ is the unscaled correlation function.

3.4.1.6 *Architectures of applied CNN models*

The architecture of a CNN typically reflects the complexity of the problem that is being solved and how much training data is available, as a complex model has more parameters than a simple model and is therefore more likely to overfit. However, much of this may be solved using proper regularization. Another restriction is the resources required to train a model; deep and complex models require far more operations and will therefore take longer to train and operate. In this study, no exhaustive hyper-parameter optimization was carried out. The applied architectures were chosen on the basis of other published models [140]. Since the models utilized three separate modalities (EEG, EOG and EMG), three separate sub-networks were constructed. These were followed by fully connected layers combining the inputs from each subnetwork, which were passed onto a softmax output, as shown in Figure 3.9b and Figure 3.11. Models that utilize memory have fully connected hidden units replaced with LSTM cells and recurrent connections added between successive segments. Networks of two different sizes are evaluated to quantify the effect of increasing complexity.

3.4.1.7 *Training of CNN models*

Training the models involves optimizing parameters to minimize a loss function evaluated across a training dataset. The loss function was defined as the cross-entropy with ℓ_2 regularization:

$$\begin{aligned} L(\boldsymbol{\omega}) &= \frac{1}{N}\sum_{i=1}^{N} H(\mathbf{y}_i, \hat{\mathbf{y}}_i) + \ell_2 \\ &= \frac{1}{N}{}_{i=1}^{N} \mathbf{y}_i \log(\hat{\mathbf{y}}_i) + (1 - \mathbf{y}_i)\log(1 - \hat{\mathbf{y}}_i) + \lambda\|\boldsymbol{\omega}\|_2^2, \end{aligned} \tag{3.23}$$

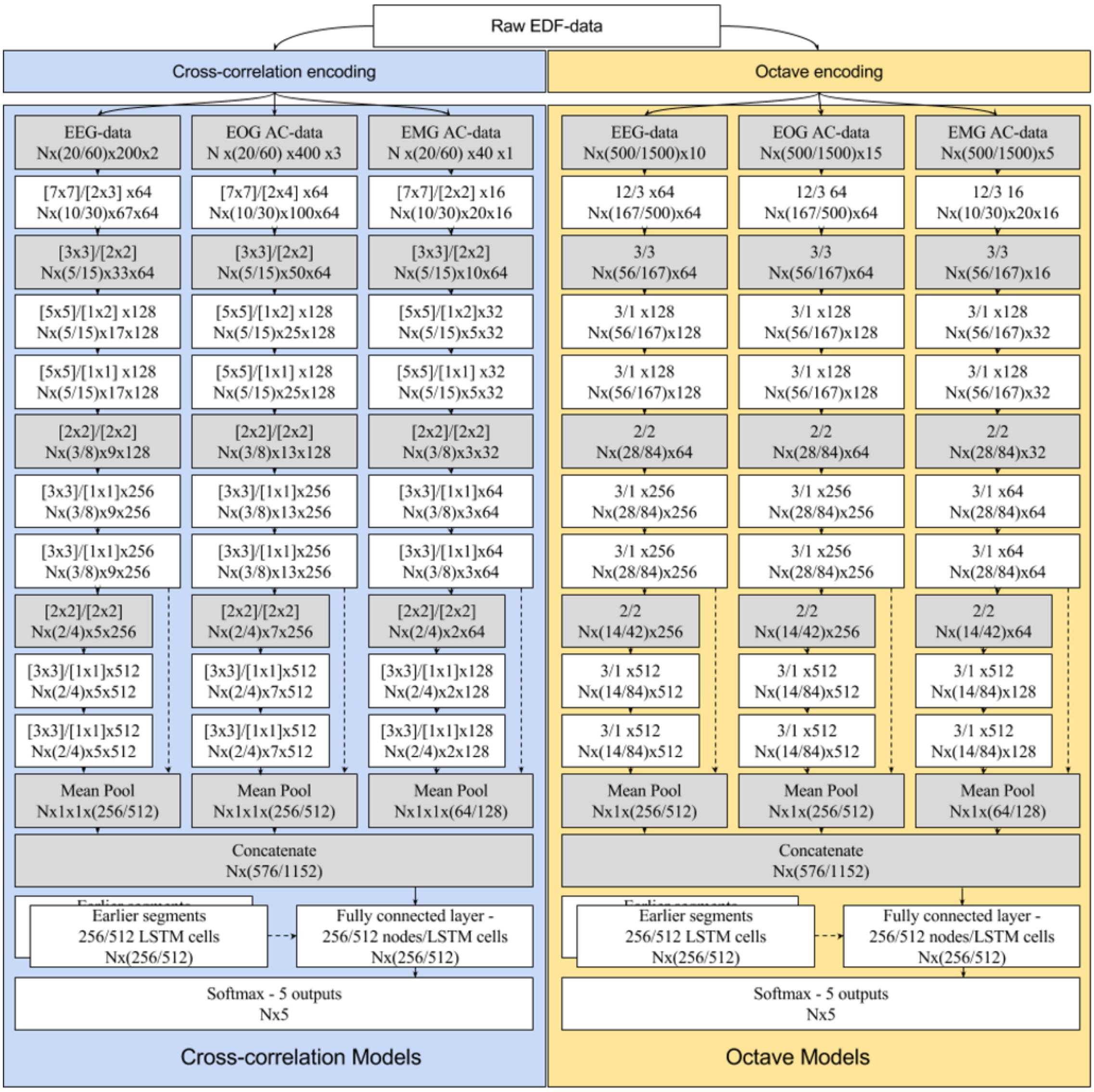

Raw EDF-data
Cross-correlation encoding
Octave encoding
EEG-data Nx(20/60)x200x2
EOG AC-data N x(20/60) x400 x3
EMG AC-data N x(20/60) x40 x1
[7x7]/[2x3] x64 Nx(10/30)x67x64
[7x7]/[2x4] x64 Nx(10/30)x100x64
[7x7]/[2x2] x16 Nx(10/30)x20x16
[3x3]/[2x2] Nx(5/15)x33x64
[3x3]/[2x2] Nx(5/15)x50x64
[3x3]/[2x2] Nx(5/15)x10x64
[5x5]/[1x2] x128 Nx(5/15)x17x128
[5x5]/[1x2] x128 Nx(5/15)x25x128
[5x5]/[1x2]x32 Nx(5/15)x5x32
[5x5]/[1x1] x128 Nx(5/15)x17x128
[5x5]/[1x1] x128 Nx(5/15)x25x128
[5x5]/[1x1] x32 Nx(5/15)x5x32
[2x2]/[2x2] Nx(3/8)x9x128
[2x2]/[2x2] Nx(3/8)x13x128
[2x2]/[2x2] Nx(3/8)x3x32
[3x3]/[1x1]x256 Nx(3/8)x9x256
[3x3]/[1x1]x256 Nx(3/8)x13x256
[3x3]/[1x1]x64 Nx(3/8)x3x64
[3x3]/[1x1]x256 Nx(3/8)x9x256
[3x3]/[1x1]x256 Nx(3/8)x13x256
[3x3]/[1x1]x64 Nx(3/8)x3x64
[2x2]/[2x2] Nx(2/4)x5x256
[2x2]/[2x2] Nx(2/4)x7x256
[2x2]/[2x2] Nx(2/4)x2x64
[3x3]/[1x1]x512 Nx(2/4)x5x512
[3x3]/[1x1]x512 Nx(2/4)x7x512
[3x3]/[1x1]x128 Nx(2/4)x2x128
[3x3]/[1x1]x512 Nx(2/4)x5x512
[3x3]/[1x1]x512 Nx(2/4)x7x512
[3x3]/[1x1]x128 Nx(2/4)x2x128
Mean Pool Nx1x1x(256/512)
Mean Pool Nx1x1x(256/512)
Mean Pool Nx1x1x(64/128)
Concatenate Nx(576/1152)
Earlier segments 256/512 LSTM cells Nx(256/512)
Fully connected layer - 256/512 nodes/LSTM cells Nx(256/512)
Softmax - 5 outputs Nx5
Cross-correlation Models
EEG-data Nx(500/1500)x10
EOG AC-data Nx(500/1500)x15
EMG AC-data Nx(500/1500)x5
12/3 x64 Nx(167/500)x64
12/3 64 Nx(167/500)x64
12/3 16 Nx(10/30)x20x16
3/3 Nx(56/167)x64
3/3 Nx(56/167)x64
3/3 Nx(56/167)x16
3/1 x128 Nx(56/167)x128
3/1 x128 Nx(56/167)x128
3/1 x128 Nx(56/167)x32
3/1 x128 Nx(56/167)x128
3/1 x128 Nx(56/167)x128
3/1 x128 Nx(56/167)x32
2/2 Nx(28/84)x64
2/2 Nx(28/84)x64
2/2 Nx(28/84)x32
3/1 x256 Nx(28/84)x256
3/1 x256 Nx(28/84)x256
3/1 x64 Nx(28/84)x64
3/1 x256 Nx(28/84)x256
3/1 x256 Nx(28/84)x256
3/1 x64 Nx(28/84)x64
2/2 Nx(14/42)x256
2/2 Nx(14/42)x256
2/2 Nx(14/42)x64
3/1 x512 Nx(14/84)x512
3/1 x512 Nx(14/84)x512
3/1 x512 Nx(14/84)x128
3/1 x512 Nx(14/84)x512
3/1 x512 Nx(14/84)x512
3/1 x512 Nx(14/84)x128
Mean Pool Nx1x(256/512)
Mean Pool Nx1x(256/512)
Mean Pool Nx1x(64/128)
Concatenate Nx(576/1152)
Earlier segments 256/512 LSTM cells Nx(256/512)
Fully connected layer - 256/512 nodes/LSTM cells Nx(256/512)
Softmax - 5 outputs Nx5
Octave Models

Figure 3.11: Specifications of each network configuration. Each block represents an operation; white blocks require multiplications and additions, whereas grey blocks are pooling or concatenations, default being maximum pooling. The top row of each block describes the size of the window and its stride, and the bottom row describes the size of the output. In this output, N is the length of a sequence, the second dimension is the segment length, and if a fourth dimension is present (CC models), the third dimension originally represents the size of the correlation function. The last dimension is the number of features in that layer. Models with a low complexity skip the third max pooling block, and go straight to mean pooling.

Table 3.12: Experimental configurations for single models

Level	Memory	Segment duration	Complexity	Encoding	Realizations
1	Simple FF	5 s	Low	Octave	1
2	LSTM	15 s	High	CC	2

where $\mathbf{y}_i$ is the true class label of the *i*th window, $\hat{\mathbf{y}}_i$ is the estimated probability of the *i*th window, ω is the parameter to be updated, and λ is the weight decay parameter set at 10^{-5}. The model parameters were initialized with $\mathcal{N}(0, 0.01)$, and trained until convergence using stochastic gradient decent with momentum [141]. Weight updates were computed as:

$$\omega_{t+1} = \omega_t + \eta \mathbf{v}_{t+1} \tag{3.24}$$

$$\mathbf{v}_{t+1} = \alpha \mathbf{v}_t - \frac{\partial \mathbf{L}}{\partial \omega_t}, \tag{3.25}$$

where α is the momentum set at 0.9, $\mathbf{v}_t$ is the learning velocity initialized at 0, and η is the learning rate, initially set at 0.005. The learning rate was gradually reduced with an exponential decay

$$\eta = \eta_0 \exp(-t/\tau), \tag{3.26}$$

where t is the number of updates and τ is a time constant, here set to 12 000.

Over-fitting was avoided using a number of regularization techniques, including batch normalization [116], weight decay [142], and early stopping [130]. Early stopping is accomplished by scheduling validation after every 50th training batch. This is done by setting aside 10% of the training data. Training is stopped if the validation accuracy starts to decrease as a sign of over-fitting. For LSTM networks, dropout set at 0.50 was included while training [143]. This ensured that model parameters generalized to the validation data and beyond. Given the stochastic nature of the training procedure, it was likely that two realizations of the same model would not lead to the same results, since models end up in different local minima. To measure the effect of this, two realizations were made of each model.

Apart from model realizations, we also investigated the effect of ensembling our sleep stage classification model. In general, ensemble models can yield higher predictive performance than any single model by attacking a classification or regression problem from multiple angles. For our specific use case, this resolves into forming a sleep stage prediction based on the predictions of all the models in the given ensemble. We tested several ensembles containing various numbers of model architectures and data encodings, as described in Tables 3.12 and 3.13.

3.4.1.8 *Performance comparisons of generated CNN models*

As stated, the influences of many different factors were analyzed. These included: using octave or CC encoding, short (5 s) or long (15 s) segment lengths, low or high complexity, with or without LSTM, and using only a single or two realizations of a given model. To quantify the effect of each in a principled manner, a 2^5-factorial experiment was designed leading to 32 different models as detailed in Tables 3.12 and 3.13. Comparisons between models was done on a per epoch basis.

Table 3.13: Experimental configurations for ensemble models

Configuration	Oct FF	Oct LSTM	CC FF	CC LSTM	FF	LSTM	Oct	CC	All models
Number of models	8	8	8	8	16	16	16	16	32

Table 3.14: Individual and overall scorer performance compared to model performance on IS-RC data.

	Overall	**Scorer 1**	**Scorer 2**	**Scorer 3**	**Scorer 4**	**Scorer 5**	**Scorer 6**
Accuracy, %							
Biased	81.3 ± 3.0	82.4 ± 6.1	84.6 ± 5.5	74.1 ± 7.9	85.4 ± 5.7	83.1 ± 9.4	78.3 ± 8.9
Unbiased	76.0 ± 3.2	77.3 ± 6.3	79.1 ± 6.3	69.0 ± 8.0	79.7 ± 6.5	77.8 ± 9.6	72.9 ± 9.2
Model, %	-	85.1 ± 4.9	83.8 ± 5.0	86.5 ± 4.3	84.3 ± 4.7	85.6 ± 4.7	87.0 ± 4.5
p-value	-	3.8×10^{-14}	7.5×10^{-9}	6.0×10^{-28}	4.7×10^{-9}	1.7×10^{-8}	7.5×10^{-19}
Cohen's κ							
Biased	61.0 ± 6.8	63.6 ± 12.2	68.4 ± 10.5	45.6 ± 19.7	69.6 ± 13.2	64.5 ± 20.9	54.5 ± 19.8
Unbiased	57.7 ± 6.1	61.3 ± 11.2	64.6 ± 10.3	43.5 ± 19.2	64.6 ± 13.1	60.9 ± 16.9	51.6 ± 16.7
Model	-	74.3 ± 12.3	72.4 ± 12.1	76.0 ± 11.8	72.7 ± 12.0	74.7 ± 12.1	76.6 ± 12.2
p-value	-	4.6×10^{-14}	7.9×10^{-10}	7.0×10^{-24}	6.4×10^{-9}	9.2×10^{-10}	2.0×10^{-20}

Both accuracy and Cohen's κ are presented with (biased) and without (unbiased) the assessed scorer included in the consensus standard in a leave-one-out fashion. Accuracy is expressed in percent, and Cohen's κ is a ratio and therefore unitless. *p*-values correspond to the paired *t*-tests between the unbiased predictions for each scorer against the model predictions on the same consensus.

3.4.2 *Results*

3.4.2.1 *Inter-scorer reliability cohort*

We assessed inter-scorer reliability using the IS-RC, a cohort of 70 PSGs scored by 6 scorers across three locations in the USA [126]. Table 3.14 displays individual scorer performance as well as the averaged performance across scorers, with top and bottom of table showing accuracies and Cohen's κ, respectively. The results are shown for each individual scorer when compared to the consensus of all scorers (biased), and compared to the consensus of the remaining scorers (unbiased). In the event of no majority vote for an epoch, the epoch was counted equally in all classes in which there was disagreement. Also shown in Table 3.14 is the model performance on the same consensus scorings as each individual scorer along with the *t*-statistic and associated *p*-value for each paired *t*-test between the model performance and individual scorer performance. At a significance level of $\alpha = 0.05$, the model performs statistically better than any individual scorer both in terms of accuracy and Cohen's κ.

Table 3.15 displays the confusion matrix for every epoch of every scorer of the inter-scorer reliability data, both unadjusted (top) and adjusted (bottom). As in [26], the biggest discrepancies occur between N1 and W, N1 and N2, and N2 and N3, with some errors also occurring between N1 and REM, and N2 and REM.

For future analyses of the IS-RC in combination with other cohorts that have been scored only by one scorer, a final hypnogram consensus was

Table 3.15: IS-RC scorer assesment.

		Concensus					
		W	N1	N2	N3	REM	Pr
Individual scorers	W	13.28 %	1.04%	0.86%	0.08%	0.23%	0.86
		13.25 %	0.98%	0.87%	0.08%	0.22%	0.86
	N1	0.79%	3.36 %	1.23%	0.03%	0.29%	0.59
		0.88%	3.61 %	1.42%	0.03%	0.31%	0.58
	N2	0.87%	2.46%	44.66 %	4.89%	0.85%	0.83
		0.84%	2.30%	45.48 %	5.92%	0.84%	0.82
	N3	0.05%	0.02%	2.58%	6.45 %	0.00%	0.71
		0.05%	0.02%	1.54%	5.41 %	0.00%	0.77
	REM	0.32%	1.00%	1.14%	0.03%	13.46 %	0.84
		0.31%	0.97%	1.16%	0.04%	13.46 %	0.84
	Se	0.87	0.43	0.88	0.56	0.91	0.81
		0.86	0.46	0.9	0.47	0.91	0.81

W: wakefulness; N1: non-rapid eye movement stage 1; N2: non-rapid eye movement stage 2; N3: non-rapid eye movement stage 3; REM: rapid eye movement; Pr, precision; Se, sensitivity.

Table 3.16: Performance of best models on various datasets compared to the six-scorer consensus. All comparisons are on a by-epoch-basis.

Test data	**Best single model**	**Accuracy, %**	**Best ensemble**	**Accuracy, %**
WSC	CC/SH/LS/LSTM/2	86.0 ± 5.0	All CC	86.4 ± 5.2
SSC+KHC				
÷ narcolepsy	CC/LH/SS/LSTM	76.9 ± 11.1	All CC	77.0 ± 11.9
+ narcolepsy	CC/LH/SS/LSTM	68.8 ± 11.0	All CC	68.4 ± 12.2
IS-RC	CC/LH/LS/LSTM/2	84.6 ± 4.6	All Models	86.8 ± 4.3

built for this cohort based on the majority vote weighted by the degree of consensus from each voter, expressed as

$$\text{Cohen's } \kappa = 1 + \frac{1 - p_o}{1 - p_e}, \tag{3.27}$$

where p_e is the baseline accuracy and p_o is the scorer accuracy, such that

$$\mathbf{y} = \arg\max \frac{\sum_{i=1}^{6} \hat{\mathbf{y}}_i \cdot \kappa_i}{\sum_{i=1}^{6} \kappa_i} \tag{3.28}$$

In this implementation, scorers with a higher consensus with the group are considered more reliable and have their assessments weighted heavier than the rest. This also avoided split decisions on end-results.

3.4.2.2 *Optimizing machine learning performance for sleep staging*

We next explored how various machine learning algorithms (see Methods) performed depending on cohort, memory (i.e., feed forward (FF) versus LSTM networks), signal segment length (short segments of 5 s (SS) versus

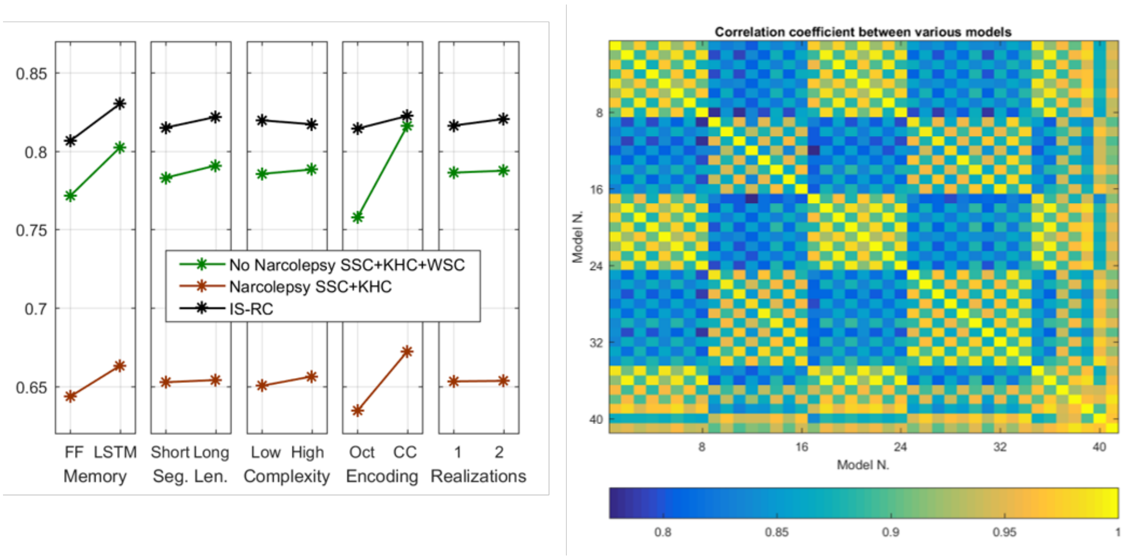


Figure 3.12: Comparisons of machine learning models. Left: Comparisons of the effect on accuracy by each factor at different settings on IS-RC data, SSC and KHC narcolepsy subjects, and the remaining SSC, KHC and WSC subjects used for testing. Right: Correlation matrix showing similarities in different model predictions, where 0 means signals are independent, and 1 means signals are completely correlated. Models 1-32 are single models, and 33-41 are ensembles. The models vary on 5 parameters, each at two levels, in the following order: Memory – FF or LSTM(1), segment size – 5 s or 15 s (2), complexity – high or low (3), encoding – CC or octave (4), realizations – 1 or 2 (5). Ensembles: All FF octave models (33), all LSTM octave models (34), all FF CC models (35), all LSTM CC models (36), all FF models (37), all LSTM models (38), all CC models (39), all octave models (40), all models (41).

long segments of 15 s (LS)), complexity (i.e., low (SH) vs. high (LH)), encoding (i.e., octave versus cross-correlation (CC) encoding, and realization type (repeated training sessions). The performance of these machine learning algorithms was compared with the six-scorer consensus in the IS-RC and with single scorer data in 3 other cohorts, the Stanford Sleep Cohort (SSC) [111], [112], the Wisconsin Sleep Cohort (WSC) [112], [125] and the Korean Hypersomnia Cohort (KHC) [111], [127] (see Datasets section in Methods for description of each cohort).

Model accuracy varies across datasets, reflecting the fact scorer performance may be different across sites, and because unusual subjects such as those with specific pathologies can be more difficult to score—a problem affecting both human and machine scoring. In this study, the worst performance was seen in the KHC and SSC with narcolepsy, and the best performance was achieved on IS-RC data (Figure 3.12a, Table 3.16). The SSC+KHC cohorts mainly contain patients with more fragmented sleeping patterns, which would explain a reduced performance. The IS-RC has the most accurate label, minimizing the effects of erroneous scoring, which therefore leads to an increased performance. Incorporating large ensembles of different models increased mean performances only slightly. (Table 3.16).

The two most important factors that increased prediction accuracy were encoding and memory, while segment length, complexity and number of realizations were less important (Figure 3.12). The effect of encoding was less prominent in the IS-RC. Prominent factor interactions include: (i) CC encoding models improve with higher complexity, whereas octave encoding models worsen; (ii) increasing segment length positively affects models with low complexity, but does not affect models with a high complexity; and (iii)

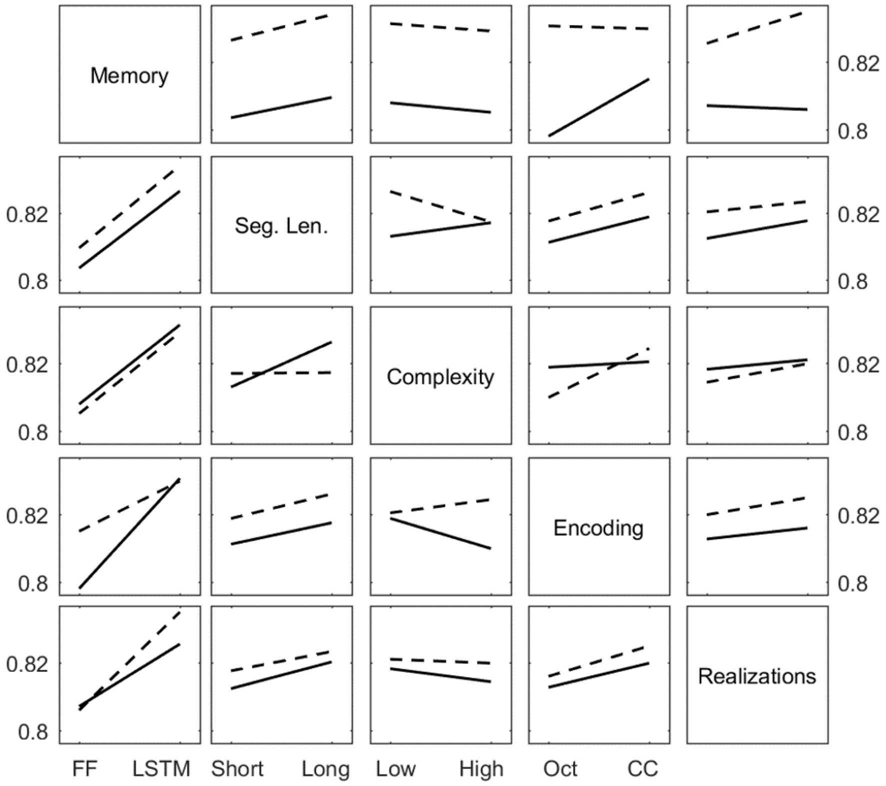


Figure 3.13: Interaction of different factors. The IS-RC data was used for this analysis. The solid and dashed lines indicate factors along the rows on levels 1 and 2, respectively.

adding memory improves models with an octave encoding more than models with a CC encoding. Because the ISRC data are considered the most reliable, we decided to use these data as benchmark for model comparison. This standard improved as more scorers were added, and the model performance increased (Figure 3.16a). The different model configurations described in this section do not represent exhaustive configuration search, and future work experiments might result in improved results.

Figure 3.14 displays typical scoring outputs (bottom panels) obtained with a single sleep study of the IS-RC cohort in comparison to 6 scorer consensus (top panel). The model results are displayed as hypnodensity graphs, representing not only discrete sleep stage outputs, but also the probability of occurrence of each sleep state for each epoch (see definition in Data labels, scoring and fuzzy logic section). As can be seen, all models performed well, and segments of the sleep study with the lowest scorer consensus (top) are paralleled by similar sleep stage probability uncertainty, with performance closest to scoring consensus achieved by an ensemble model described below (second to top).

3.4.2.3 *Final implementation of automatic sleep scoring algorithm*

Because of model noise, potential inaccuracies and the desire to quantify uncertainty, the final implementation of our sleep scoring algorithm is an ensemble of different CC models with small variations in model parameters, such as the number of feature-maps and hidden nodes. This was achieved

Table 3.17: Confusion matrix displaying the relation between different targets and the ensemble estimate.

		Target					
		W	N1	N2	N3	REM	Pr
Model prediction	W	14.08%	0.35%	0.88%	0.01%	0.08%	0.91
		16.68%	0.15%	0.44%	0.00%	0.02%	0.96
	N1	1.13%	1.78%	3.00%	0.00%	0.36%	0.28
		0.47%	0.88%	1.15%	0 %	0.12%	0.34
	N2	0.29%	0.59%	52.58%	1.27%	0.66%	0.95
		0.12%	0.25%	56.30%	0.34%	0.32%	0.98
	N3	0.00%	0 %	2.13%	4.87%	0 %	0.7
		0 %	0 %	1.09%	4.23%	0 %	0.91
	REM	0.54%	1.17%	0.78%	0 %	13.45%	0.84
		0.40%	0.73%	0.41%	0 %	15.86%	0.91
	Se	0.88	0.46	0.89	0.79	0.92	0.87
		0.94	0.44	0.95	0.92	0.97	0.94

Top row: unweighted consensus. Bottom row: weighted by the scorer agreement at each epoch. The number of analyzed epochs were 53 009 (un-weighted) and 36 032 (weighted). W: wakefulness; N1: non-rapid eye movement stage 1; N2: non-rapid eye movement stage 2; N3: non-rapid eye movement stage 3; REM: rapid eye movement; Pr, precision; Se, sensitivity.

by randomly varying the parameters between 50 and 150% of the original values using the CC/SH/LS/LSTM as a template (this model achieved similar performance to the CC/LH/LS/LSTM while requiring significantly less computational power).

We trained 16 models, and at each segment (5 s, 10 s, 15 s and 30 s) the mean and variance of model estimates were calculated. As expected, the relative model variance (standardized to the average variance in a correct wakefulness prediction) is generally lower in correct predictions and this can be used to inform users about uncertain/incorrect estimates. To demonstrate the effectiveness of this final implementation, the average of the models is shown alongside the distribution of 5234 ± 14 scorers on 150 epochs, a dataset provided by the AASM (AASM inter-scorer reliability (ISR) dataset, (see Datasets section in Methods). On these epochs, the AASM ISR achieved a 90% agreement between scorers. In comparison, the model estimates reached a 95% accuracy compared to the AASM consensus (Fig. 2b). Using the model ensemble and reporting on sleep stage probabilities and inter-model variance for quality purpose constitute the core of our sleep scoring algorithm.

3.4.2.4 *Ensemble/best model performance*

Table 3.15 reports on concordance for our best model, the ensemble of all CC models. Concordance is presented in a weighted and unweighted manner, between the best model estimate and scorer consensus (Table 3.17). Weighing of a segment was based on scorer confidence and serves to weigh down

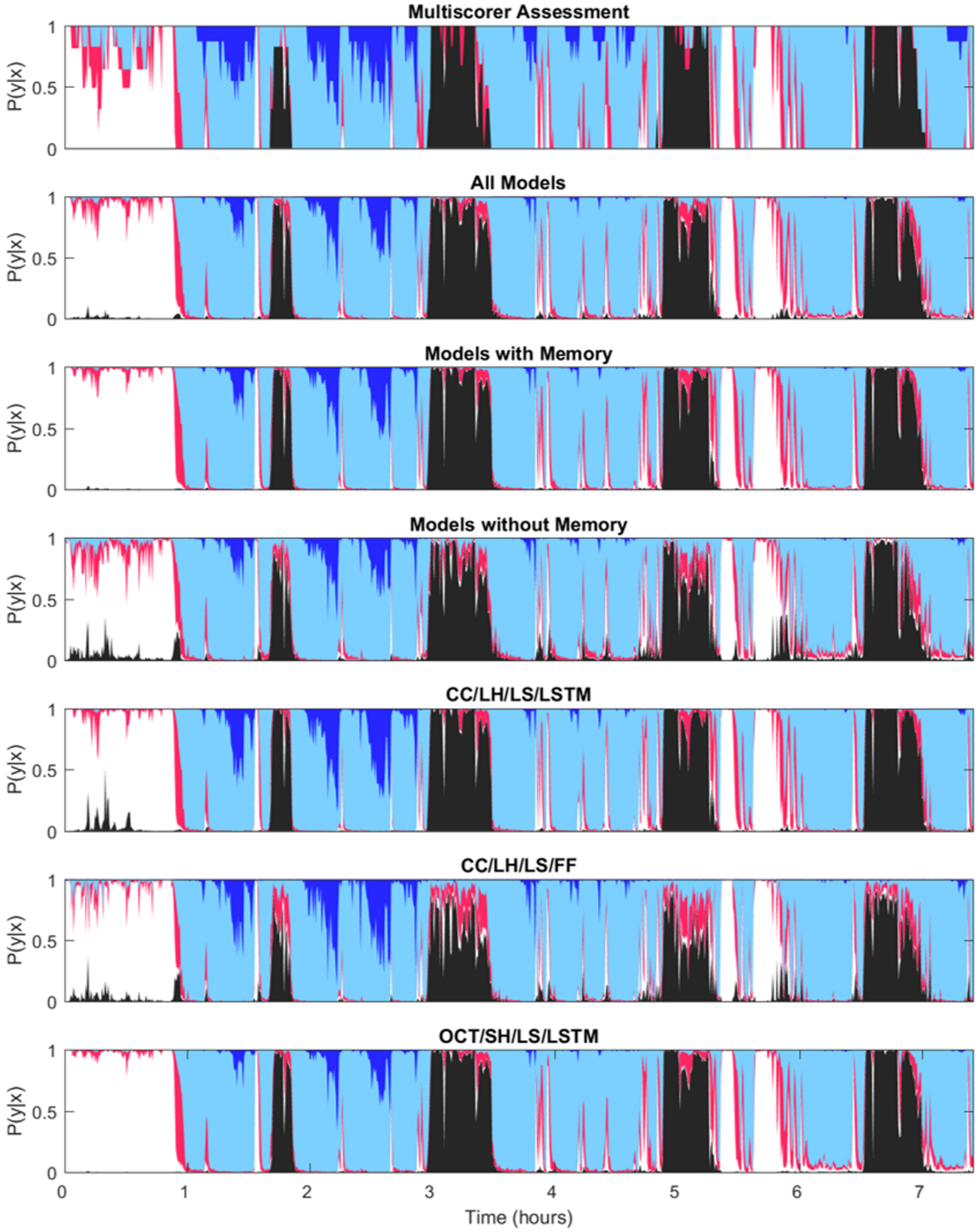


Figure 3.14: The figure displays the hypnodensity graph. Displayed models are in order: multiple scorer assessment (1); ensembles: All models, those with memory (LSTM) and those without memory (FF) (2–4); single models. OCT is octave encoding, Color codes: white, wake; red, N1; light blue, N2; dark blue, N3; black, REM.

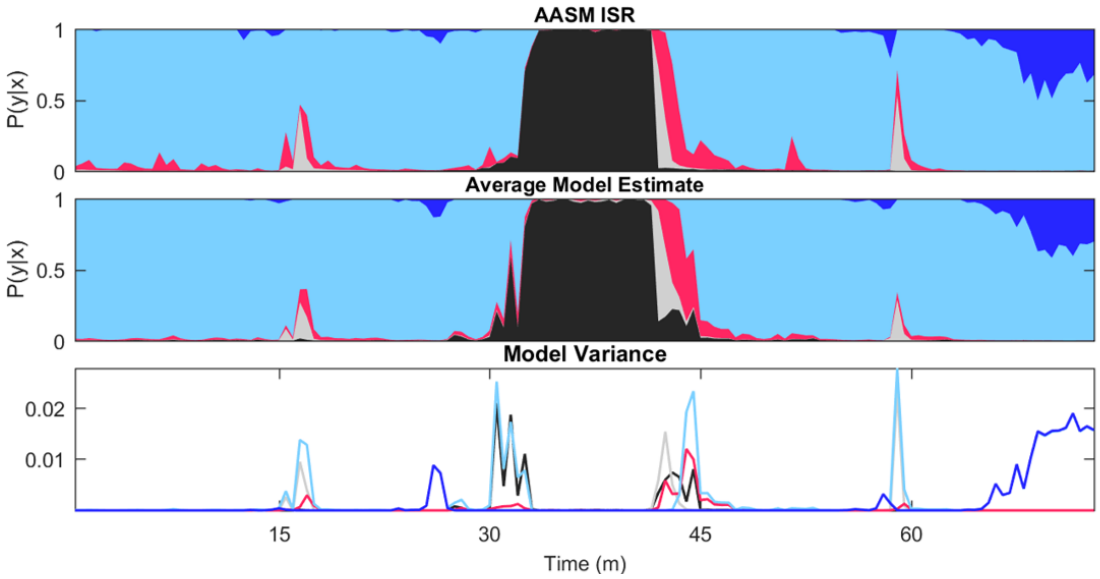
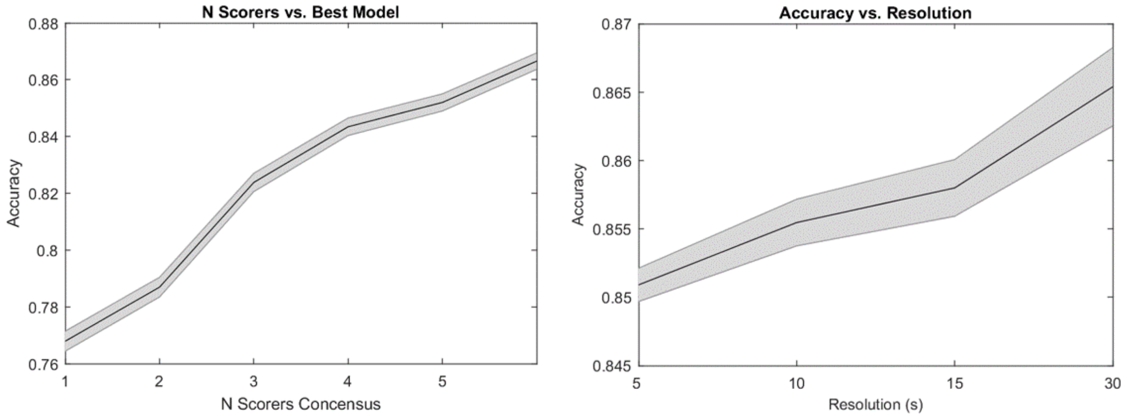


Figure 3.15: The 150 epochs of a recording from the AASM Inter-Scorer Reliability program (ISR) are analyzed by 16 models with randomly varying parameters, using the CC/SH/LS/LSTM model as a template. These data were also evaluated by 5234 ± 14 different scorers. The distribution of these is shown on top, the average model predictions are shown in the middle, and the model variance is shown at the bottom.

controversial segments. For each recording i, the epoch-specific weight w_n and weighted accuracy αw were calculated as:

$$\begin{aligned} w_n &= \max_{z \in \mathcal{Z}}(\mathbf{P}(\mathbf{y}_n \mid \mathbf{x}_n)) - \ell^2_{\mathcal{Z}}(\mathbf{P}(\mathbf{y}_n \mid \mathbf{x}_n)), \\ \alpha^{(i)}_w &= \frac{1}{\sum_n w_n} \sum_n w_n \left(\arg\max_{m \in \mathcal{M}}(\mathbf{P}_m(\hat{\mathbf{y}}_n|\mathbf{x}_n)) \cap \arg\max_{z \in \mathcal{Z}}(\mathbf{P}_z(\mathbf{y}_n|\mathbf{x}_n)) \right), \end{aligned} \tag{3.29}$$

where $\ell^2_{\mathcal{Z}}(\mathbf{P}(\mathbf{y}_n \mid \mathbf{x}_n))$ is the second most likely stage assessed by the set of scorers (experts) denoted by $\mathcal{Z}$, of the nth epoch in a sleep recording. As with scorers, the biggest discrepancies occurred between wake versus N1, N1 versus N2 and N2 versus N3. Additionally, the weighted performance was almost universally better than the unweighted performance, raising overall accuracy from 87 to 94%, indicating a high consensus between automatic scoring and scorers in places with high scorer confidence. An explanation for these results could be that both scorers and model are forced to make a choice between two stages when data are ambiguous. An example of this may be seen in Fig. 2a. Between 1 and 3 h, several bouts of N3 occur, although they often do not reach the threshold for being the most likely stage As time progresses, more evidence for N3 appears reflecting increased proportion of slow waves per epoch, and confidence increases, which finally yields "definitive" N3. This is seen in both model and scorer estimates. Choosing to present the data as hypnodensity graphs mitigates this problem. The various model estimates produce similar results, which also resemble the scorer assessment distribution, although models without memory fluctuate slightly more, and tend to place a higher probability on REM sleep in periods of wakefulness, since no contextual information is provided



(a)

(b)

Figure 3.16: Accuracy per scorer and by time resolution. (a) The effect on scoring accuracy as golden standard is improved. Every combination of N scorers is evaluated in an unweighted manner and the mean is calculated. Accuracy is shown with mean (solid black line) and a 95% confidence interval (gray area). (b) Predictive performance of best model at different resolutions. Performance is shown as mean accuracy (solid black line) with a 95% confidence interval (gray area).

3.4.2.5 *Influences of sleep pathologies*

To see how much may be attributed to various pathologies, five different analyses of variance were made, with accuracy as the dependent variable, using cohort, age (grouped as age < 30, $30 \leqslant$ age < 50, and age $\geqslant 50$) and sex as covariates, investigating the effect of insomnia, OSA, restless leg syndrome (RLS), periodic leg movement index (PLMI) and NT1 on accuracy of our machine learning routine versus human scoring. This was performed in the cohort mentioned above with addition of the Austrian Hypersomnia Cohort (AHC). The *p*-values obtained from paired *t*-testing for each condition were 0.75 (insomnia), 7.53×10^{-4} (OSA), 0.13 (RLS), 0.22 (PLMI) and 1.77×10^{-15} (NT1) respectively, indicating that only narcolepsy had a strong effect on scorer performance. Additionally, in the context of narcolepsy, cohort and age yielded *p*-values between 3.69×10^{-21} and 2.81×10^{-82} and between 0.62 and 6.73×10^{-6}, respectively. No significant effect of gender was ever noted. Cohort effects were expected and likely reflect local scorer performances and differences in PSG hardware and filter setups at every site. Decreased performance with age likely reflects decreased EEG amplitude, notably in N3/slow wave sleep amplitude with age36.

3.4.2.6 *Resolution of sleep stage scoring*

Epochs are evaluated with a resolution of 30 s, a historical standard that is not founded in anything physiological, and limits the analytical possibilities of a hypnogram. Consequently, it was examined to what extent the performance would change as a function of smaller resolution. Only the models using a segment size of 5 s were considered. Segments were averaged to achieve performances at 5, 10, 15 and 30 s resolutions, and the resulting performances in terms of accuracy are shown in Figure 3.16b. Although the highest performance was found using a resolution of 30 s, performance dropped only slightly with decreasing window sizes.

3.4.3 *Discussion*

In recent years, machine learning has been used to solve similar or more complex problems, such as labeling images, understanding speech and translating language, and have seen advancement to the point where humans are now sometimes outperformed [95], [98], [144], while also showing promising results in various medical fields [145]–[150]. Automatic classification of sleep stages using automatic algorithms is not novel [78], [151], but only recently has this type of machine learning been applied and the effectiveness has only been demonstrated in a small numbers of sleep studies [36], [39], [152]–[154]. Because PSGs contain large amounts of manually annotated gold standard data, we hypothesized this method would be ideal to automatize sleep scoring. We have shown that machine learning can be used to score sleep stages in PSGs with high accuracy in multiple physical locations in various recording environments, using different protocols and hardware/software configurations, and in subjects with and without various sleep disorders.

After testing various machine learning algorithms with and without memory and specific encodings, we found increased robustness using a consensus of multiple algorithms in our prediction. The main reason for this is likely the sensitivity of each algorithm to particular aspects of each individual recording, resulting in increased or decreased predictability. Figure 3.12b displays the correlations between different models. Those incorporating an ensemble of different models generally have a higher overall correlation coefficient than single models, and since individual models achieve similar performances, it stands to reason that these would achieve the highest performance.

In addition to the stochastic nature of the training, one potential source for this variability was that recordings were conducted in different laboratories that were using different hardware and filters, and had PSGs scored by technicians of various skill levels. Another contributor was the presence of sleep pathologies in the dataset that could influence machine learning. However, of the pathologies tested, only narcolepsy had a very significant effect on the correspondence between manual and machine learning methods. This was not surprising as the pathology is characterized by unusual sleep stage transitions, for example, transitions from wake to REM sleep, which may make human or machine learning staging more difficult. This result suggests that reporting inter-model variations in accuracy for each specific patient has value in flagging unusual sleep pathologies.

Unlike previous attempts using automatic detector validations, we were able to include 70 subjects scored by 6 technicians in different laboratories from the IS-RC to independently validate our best automatic scoring consensus algorithm [126]. This allowed us to estimate the performance at 87% in comparison to the performance of a consensus score for every epoch among six expert technicians, see Table 3.14. Including more scorers produces a better gold standard, and as Figure 3.16a indicates, the model accuracy also increases with more scorers. Naturally, extrapolating from this should be done with caution; however, it is reasonable to assume that the accuracy would continue to increase with increased scorers. In comparison, performance of any individual scorer ranges from 74 to 85% when compared to the same six-scorer gold standard, keeping in mind this performance is artificially inflated since the same scorers evaluated are included in the gold standard. The best model achieves 87% accuracy using 5 scorers and

The unbiased performance of any scorer versus consensus of remaining 5 scorers range from 69 % to 80 %.

achieves statistically significantly higher performance values than all scorers, as shown in Figure 3.16a and Table 3.14.

As with human scorers, the biggest discrepancies in machine learning determination of sleep stages occurred between wake versus N1, N1 versus N2 and N2 versus N3. This is logical as these particular sleep stage transitions are part of a continuum, artificially defined and subjective. To give an example: an epoch comprised of 18% slow wave activity is considered N2 while an epoch comprised of 20% slow wave activity qualifies as N3 according to AASM guidelines [43]. Overall, data indicate that our machine learning algorithm performs better than individual scorers, as typically used in clinical practice, or similar to the best of 5 scorers in comparison to a combination of 5 experts scoring each epoch by consensus. It is also able to score at higher resolution, i.e., 5 s, making it unnecessary to score sleep stages by 30 s epochs, an outdated rule dating from the time sleep was scored on paper.

In conclusion, models which classify sleep by assigning a membership function to each of five different stages of sleep for each analyzed segment were produced, and factors contributing to the performance were analyzed. The models were evaluated on different cohorts, one of which contained 70 subjects scored by 6 different sleep scoring technicians, allowing for inter-scorer reliability assessments. The most successful model, consisting of an ensemble of different models, achieved an accuracy of 87% on this dataset, and was statistically better performing than any individual scorer. It was also able to score sleep stages with high accuracy at lower time resolution (5 s), rendering the need for scoring per 30 s epoch obsolete. When predictions were weighted by the scorer agreement, performance rose to 95%, indicating a high consensus between the model and human scorers in areas of high scorer agreement. A final implementation was made using an ensemble with small variations of the best single model. This allowed for better predictions, while also providing a measure of uncertainty in an estimate.

3.5 CHAPTER SUMMARY

RH 1: *Advanced biomedical signal processing and machine learning algorithms can be used for efficient, high-performing analysis of sleep studies with regards to sleep stages*

Sleep stage classification is performed manually by experts in sleep clinics leading to major inter- and intra-variability [24], [26], [28], [29], [77]. One potential way to overcome this challenge is to assist or augment the manual scoring with fully automatic intelligent systems (**RH 1**), that provide consistency and robustness in the analysis of sleep patterns. In this chapter, we introduced methods for automating sleep stage classification using deep neural networks with two separate model frameworks to answer research questions **RQ 1.1**, **RQ 1.2**, and **RQ 1.3**.

RQ 1.1: *can sleep stages be effectively and reliably classified using novel machine learning algorithms*

Section 3.2 described the initial version of the MASSC algorithm, and end-to-end deep learning model based on the ResNet-50 architecture. We trained and tested the algorithm on a collective of 2310 PSGs using three different training strategies, the best of which yielded a high accuracy value of 84.1% and a Cohen's κ of 0.746. In view of the large number of PSG recordings included in the study, these numbers compare favorably to the current state-of-the-art in automatic sleep stage scoring, as well as the reported inter-rater reliability measures described in Section 2.3.1.

RQ 1.2: *in cases of multiple available data sources, is it better to have more volume or more diverse data*

However, like many other published papers on automatic sleep stage classification, the results reported in Section 3.2 are based solely on a single cohort of PSGs, which immediately raises concerns over the actual generalizability of the model. In Section 3.3 we applied an updated version of the MASSC algorithm in four different experimental settings using five cohorts differing in size, demographics, inherent co-morbidities and recording setups. We found that training models on individual cohorts yielded large variations in classification performance both in LOCI and LOCO training configurations. Strikingly, we found consistently higher sleep stage classification accuracy as a function of the data fraction by mixing cohorts in the training data compared to training models on single cohorts. Using 100% of the training data, our model achieved an accuracy of 86.9% and a Cohen's κ of 0.799, which in light of the high numbers of both training and testing records compares favorably to the state-of-the-art, as well as our previous reported results in Section 3.2.

RQ 1.3: *how can we guarantee that such a system is stable with respect to the impact of sleep disorders*

The final section described the sleep stage classification part of the STAGES model. Here, we used specific transformations of the input PSG signals coupled with multiple realizations of a deep neural network architecture to create a final ensemble model for classifying sleep stages. Based on a total 2784 PSGs, the best performing model as determined by a 2^5-factorial experimental design yielded an accuracy of 86.8% on a dataset scored by six technicians, while outperforming every single one based on both a biased and unbiased consensus score. The model was also shown to be stable with respect to the presence of several sleep disorders, with the exception of narcolepsy which had a significant impact on the algorithm.

The implications of this will be detailed in Chapter 5

4 SLEEP EVENT DETECTION

What, Morty, you want me to show you my math?

— Rick Sanchez
Rick and Morty, season 1, episode 6

This chapter presents the methods developed for detection of sleep events. The multi-modal sleep event detection (MSED) algorithm for arousal and limb movement detection, originally published in [115], is presented first in Section 4.2 and followed by the updated version in Section 4.3. A method for improving single-EEG arousal detection is also included in Section 4.4. The chapter will conclude with a summary and discussion of the main findings of the individual research items in Section 4.5.

Parts of this chapter have been modified from the following original publications:

- **A. N. Olesen**, S. Chambon, V. Thorey, P. Jennum, E. Mignot, and H. B. D. Sorensen, "Towards a Flexible Deep Learning Method for Automatic Detection of Clinically Relevant Multi-Modal Events in the Polysomnogram", *2019 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Berlin, Germany: IEEE, 2019, pp. 556–561. DOI: 10.1109/EMBC.2019.8856570[1]

- **A. N. Olesen**, P. Jennum, E. Mignot, and H. B. D. Sorensen, *A multi-modal sleep event detection algorithm for clinical sleep analysis*, 2020, (*in preparation*)

- **A. N. Olesen**, P. Jennum, E. Mignot, and H. B. D. Sorensen, *Deep transfer learning for improving single-EEG arousal detection*, 2020. arXiv: 2004.05111 [cs.CV], (*accepted*, IEEE EMBC 2020)

4.1 RESEARCH BACKGROUND

As described in Chapter 2, a correct diagnosis of sleep disorders is predicated on precise scoring of sleep stages as well as accurate scoring of discrete

[1] ©2019 IEEE

sleep events. However, the current gold standard of manual analysis by experienced technicians is inherently biased and inconsistent due to low inter-rater reliability on the scoring of sleep stages [24], [26], [28], arousals [68], and respiratory events [27], as described in Section 2.3. As manual analysis of PSGs is also time-consuming and prone to scorer fatigue, there is a need for efficient systems that provide deterministic and reliable scorings of sleep studies.

Although classification of sleep stages in large cohorts has been explored with good results [73], [82], [85], [89], [157], reliable and consistent detection and classification of discrete PSG events in large cohorts remains largely unexplored. Two studies recently proposed methods for automatic detection of arousals [158], and leg movements [159], and both tested their algorithms on a subset of data from two cohorts scored by multiple technicians. Both studies found that their algorithms could score as well as, or in some cases outperform, human scorers. However, both methods predicted events at discrete intervals, which might introduce biases in the decision making of when to merge and split certain predictions.

Recent studies on certain micro-events in sleep have indicated that deep learning methods reliably detect and annotate sleep spindles and K-complexes with start time and duration [113], [114]. These studies proposed a single-shot event detection algorithm, that parallels the YOLO and SSD algorithms used for object detection in 2D images [160]–[162], but were limited in scope by detecting events only at the EEG level, and did not explicitly take advantage of the temporal connection of the detected events. Additionally, experiments were carried out on a small-scale database [113].

Designing reliable and robust systems for automated sleep analysis based on machine learning algorithms often requires multiple heterogeneous data sources of sufficient size. However, due to differences in clinical practice, few datasets in sleep science have standardized recording setups despite guidelines from the AASM. This creates a *channel mismatch problem*, in which the overlap between our source and target domains is small, and the domains are possibly disjointed. Deep transfer learning has recently been investigated to solve the channel mismatch problem when training and testing sleep stage classification models [163], [164]. By using a fine-tuning strategy the authors significantly improved the performance of sleep stage scoring models when trained on various combinations of EEG and EOG channels.

Using a pre-trained deep neural network on a separate domain.

4.1.1 *Research motivation and objectives*

Motivated by these unresolved issues in sleep scoring, we were interested in the following research questions specifically related to research hypothesis **RH 2**:

RH 2: *Advanced biomedical signal processing and machine learning algorithms can be used for efficient, high-performing analysis of sleep studies with regards to sleep events.*

RQ 2.1 can sleep events be detected precisely and reliably using novel machine learning algorithms?

RQ 2.2 can the detection of one event class modulate the detection of an event from another class?

RQ 2.3 how can we overcome the channel mismatch problem for sleep event detection?

In this case, *detection* covers both *localization* and *classification* of sleep events.

Localization places an unclassified sleep event in the time domain, while classification determines the class, or type, of the sleep event.

Derived from the research hypothesis and associated questions, the following research objectives were formulated:

(i) a single model should detect multiple sleep events independently;

(ii) the events should be annotated with a start and duration directly to avoid unnecessary postprocessing of predictions.

The following sections describe the steps taken to complete the posed design objectives and answer the research questions.

4.2 PAPER IV: TOWARDS A FLEXIBLE DEEP LEARNING METHOD FOR AUTOMATIC DETECTION OF CLINICALLY RELEVANT MULTI-MODAL EVENTS IN THE POLYSOMNOGRAM

ABSTRACT: Much attention has been given to automatic sleep staging algorithms in past years, but the detection of discrete events in sleep studies is also crucial for precise characterization of sleep patterns and possible diagnosis of sleep disorders. We propose here a deep learning model for automatic detection and annotation of arousals and leg movements. Both of these are commonly seen during normal sleep, while an excessive amount of either is linked to disrupted sleep patterns, excessive daytime sleepiness impacting quality of life, and various sleep disorders. Our model was trained on 1485 subjects and tested on 1000 separate recordings of sleep. We tested two different experimental setups and found optimal arousal detection was attained by including a recurrent neural network module in our default model with a dynamic default event window (F1 = 0.75), while optimal leg movement detection was attained using a static event window (F1 = 0.65). Our work show promise while still allowing for improvements. Specifically, future research will explore the proposed model as a general-purpose sleep analysis model.

4.2.1 *Materials & Methods*

4.2.1.1 *MrOS Sleep Study*

The MrOS Sleep Study is a part of the larger Osteoporotic Fractures in Men Study with the objective of researching the links between sleep disorders, fractures, cardiovascular disease and mortality in older males (>65 years) [103]–[105]. Between 2003 and 2005, 3135 of the original 5994 participants were recruited to undergo full-night PSG recording at six centers in the US at two separate visits (visit 1 and visit 2) with following 3 to 5-day actigraphy studies at home. The resulting PSG studies were subsequently scored by experienced sleep technicians for standard sleep variables including sleep stages, leg movements, arousals, and respiratory events.

4.2.1.2 *Included events and signals*

In this study, we only considered the detection of two PSG events: arousals and leg movements. These events are characterized by a start time and a duration, which we extracted from 2907 PSG studies from visit 1 available from the National Sleep Research Resource repository [107], [108]. From each PSG study, we extracted left and right central EEG, left and right EOG, chin EMG, and EMG from the left and right anterior tibialis. EEG and EOG channels were referenced to the contralateral mastoid process, while a leg EMG channel was synthesized by referencing left to right. Any PSG without the full set of channels or without any event scoring was eliminated from further analysis.

Table 4.1: MrOS data demographics.

	TRAIN	EVAL	TEST	p-value
N	1485	165	1000	
Age, years	76.4 ± 5.5	76.6 ± 4.9	76.4 ± 5.6	0.631
BMI, kg s^{-2}	27.2 ± 3.8	27.2 ± 3.4	27.1 ± 3.7	0.879
AHI, h^{-1}	12.8 ± 12.9	10.6 ± 11.8	11.9 ± 12.8	0.029
ArI, h^{-1}	23.6 ± 11.5	24.1 ± 12.2	23.4 ± 11.8	0.607
PLMI, h^{-1}	34.8 ± 37.0	37.8 ± 38.9	37.3 ± 38.0	0.204

Continuous variables were tested for significance with Mann-Whitney U-tests. Significant p-values at $\alpha = 0.05$ are shown in bold. BMI: body-mass index; AHI: apnea-hypopnea index; ArI: arousal index; PLMI: periodic leg movement index.

4.2.1.3 *Subset demographics and partitioning*

In total, 2650 out of the 2907 PSGs available from visit 1 were included in this study. These were partitioned into TRAIN, EVAL, and TEST sets containing 1485, 165, and 1000 studies, respectively. A subset of key demographic and PSG variables are presented in Table 4.1.

4.2.1.4 *Signal preprocessing*

All signals were resampled to $f_s = 128\,\text{Hz}$ using poly-phase filtering with a Kaiser window ($\beta = 5.0$) before subsequent filtering according to AASM criteria [43]. Briefly, EEG and EOG channels were subjected to a 4th order digital Butterworth band-pass filter with a 0.3 Hz to 35 Hz passband, while chin and leg EMG channels were filtered with a 4th order digital Butterworth high-pass filter with a 10 Hz cutoff frequency. All filters were implemented using zero-phase filtering. Lastly, each channel was normalized by subtracting the channel mean and dividing by the channel standard deviation across the entire night.

The zero-phase filtering procedure filters a signal in the forward direction, and then in the reverse direction, while matching the initial conditions of the filter in the reverse direction.

4.2.1.5 *Detection model overview*

In brief, the proposed model receives as input a tensor $\mathbf{x} \in \mathbb{R}^{C \times T}$ containing C channels of data in a segment of T samples, along with a set of events $\{\varepsilon_i \in \mathbb{R}^2 \mid \varepsilon_i = (\rho_i, \delta_i), i \in [\![N_\mathbf{x}]\!]\}$, were $N_\mathbf{x}$ is the number of events in the associated time segment and (ρ_i, δ_i) are the start time and duration of event ε_i. The objective of the deep learning model f is then to infer $\{\varepsilon_i\}$ given $\mathbf{x}$. To do this, a set of default events $\{\varepsilon_j^d \in \mathbb{R}^2 \mid j \in [\![N_d]\!], N_d = T/\tau\}$ is generated over the segment of T samples, where τ is the size of each default event window in samples. The model outputs probabilities for K classes including the default, non-event class for each default event window. The probability for a given class k in the default event window ε_j^d must be greater than a classification threshold θ_{clf}. In order to select among many possible candidates of predicted events, all predicted events of class k over the possible events in N_d is subjected to non-maximum suppression using the intersection over union (IoU) as in [160], [161]. A high-level schematic of the detection model is shown in Figure 4.1.

The intersection over union (IoU) is also known as the Jaccard index.

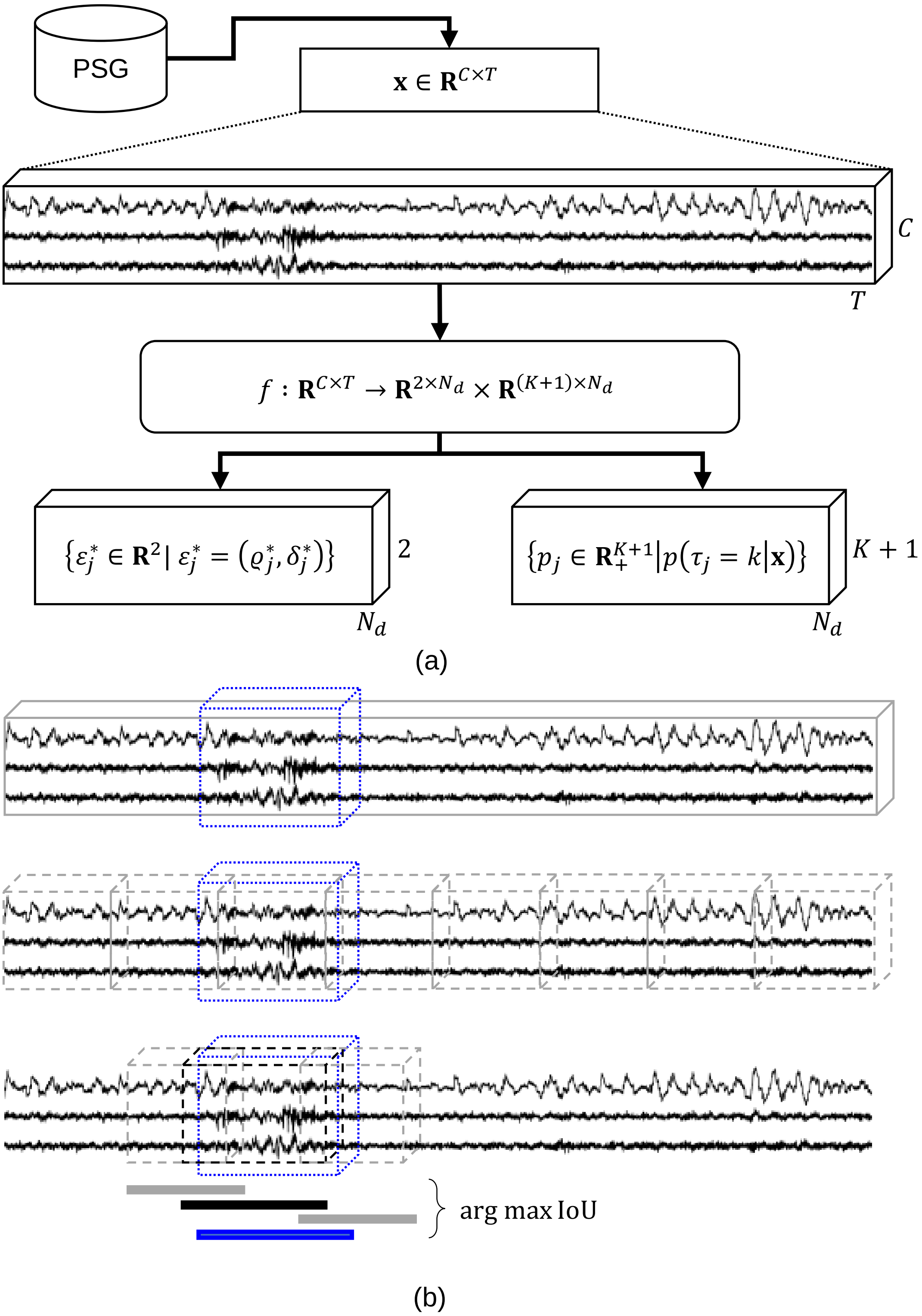

PSG
$\mathbf{x} \in \mathbf{R}^{C \times T}$
C
T
$f : \mathbf{R}^{C \times T} \to \mathbf{R}^{2 \times N_d} \times \mathbf{R}^{(K+1) \times N_d}$
$\{\varepsilon_j^* \in \mathbf{R}^2 \mid \varepsilon_j^* = (\varrho_j^*, \delta_j^*)\}$
2
N_d
$\{p_j \in \mathbf{R}_+^{K+1} | p(\tau_j = k | \mathbf{x})\}$
$K + 1$
N_d
(a)
arg max IoU
(b)

Figure 4.1: Schematic of proposed event detection procedure. **(a)** Input data $\mathbf{x}$ is fed to the model f, which outputs predictions for event classes and localizations for each default event in ε^d. **(b)** The IoU for each predicted ε_j^* is then calculated with respect to the true event ε_i and non-maximum suppression is applied to match up true events and predictions. In the current case, the predicted event marked in black has the highest IoU with the true event in blue. For more information, see [113], [114], [162].

Table 4.2: Event detection network architecture.

Module	Input dim.	Output dim.	Type	Kernel	Filters	Stride	Activation
ϕ_C	(C, T)	(C, T)	1D conv	C	C	1	linear
$\phi_{T,\text{init}}$	(C, T)	$(8, T)$	1D conv	3	8	1	–
	$(8, T)$	$(8, T)$	batch norm.	–	8	–	ReLU
	$(8, T)$	$(8, T/2)$	1D maxpool	2	–	2	–
$\phi_{T,n}$	$(2^{n+1}, T/2^{n-1})$	$(2^{n+2}, T/2^{n-1})$	1D conv	3	2^{n+2}	1	–
	$(2^{n+2}, T/2^{n-1})$	$(2^{n+2}, T/2^{n-1})$	batch norm.	–	2^{n+2}	–	ReLU
	$(2^{n+2}, T/2^{n-1})$	$(2^{n+2}, T/2^{n})$	1D maxpool	2	–	2	–
ϕ_R	$(\tilde{C}, \tilde{T})$	$(2 \times \tilde{C}, \tilde{T})$	bGRU	$\tilde{C}$	–	–	–
ψ_{clf}	$(\tilde{C}, \tilde{T})$	$(KN_d, 1)$	1D conv	$\tilde{T}$	KN_d	$\tilde{T}$	softmax over K filters
ψ_{loc}	$(\tilde{C}, \tilde{T})$	$(2N_d, 1)$	1D conv	$\tilde{T}$	$2N$	$\tilde{T}$	linear

ϕ_C, linear mixing module; ϕ_T, temporal feature extraction module; ϕ_R, recurrent neural network module; ψ_{clf}, event classification module; ψ_{loc}, event localization module; C, number of input channels; T, number of samples in segments; $\tilde{C} = 2^{2+n_{\max}}$, number of output channels; K, number of event classes; N_d, number of default events in segment; $\tilde{T} = T/2^{n_{\max}}$, reduced temporal dimension; bGRU, bidirectional gated recurrent unit; ReLU, rectified linear unit.

4.2.1.6 *Network architecture*

The architecture for the proposed PSG event detection model closely follows the event detection algorithms described in [113], [114], albeit with some specific changes. An overview of the proposed network in the model f is provided in Table 4.2. Briefly, the model comprises three modules:

1. a channel mixing module $\phi_C : \mathbb{R}^{C \times T} \to \mathbb{R}^{C \times T}$;
2. a feature extraction module $\phi_T : \mathbb{R}^{C \times T} \to \mathbb{R}^{\tilde{C} \times \tilde{T}}$;
3. and an event detection module ψ,

the latter contains two submodules performing event classification $\psi_{\text{clf}} : \mathbb{R}^{\tilde{C} \times \tilde{T}} \to \mathbb{R}^{(K+1) \times N_d}$ and event localization $\psi_{\text{loc}} : \mathbb{R}^{\tilde{C} \times \tilde{T}} \to \mathbb{R}^{2 \times N_d}$, respectively. ϕ_{clf} outputs the probability of the default, non-event class and K event classes, while ϕ_{loc} predicts a start time and a duration of all predicted events relative to a specific default event window. The channel mixing module ϕ_C receives a segment of input data $x \in \mathbb{R}^{C \times T}$, where C is the number of input channels and T is the number of time samples in the given segment, and subsequently performs linear channel mixing using 1D convolutions to synthesize C new channels. Following ϕ_C, the feature extraction module ϕ_T consists of $n_{\max}$ blocks with the first block $\phi_{T,1} : \mathbb{R}^{C \times T} \to \mathbb{R}^{8 \times T/2}$ and the nth block $\phi_{T,n} : \mathbb{R}^{2^{n+1} \times T/2^{n-1}} \to \mathbb{R}^{2^{k+2} \times T/2^{n}}$. All $n_{\max}$ blocks implement $\phi_{T,n}$ using 1D convolution layers followed by batch normalization of the feature maps, rectified linear unit activation, and final 1D maximum pooling layers across the temporal dimension. Kernel sizes and strides for convolution and max. pool. layers in ϕ_T were set to 3 and 1, and 2 and 2, respectively, while the number of feature maps in $\phi_{T,n}$ was set to 2^{n+2}. The event classification submodule ψ_{clf} is implemented a 1D convolution layer across the entire data volume using $(K+1)N_d$ feature maps of size and stride $\tilde{T} = T/2^{n_{\max}}$, where $K \in \mathbf{N}$ is the number of event classes to be detected and $N_d \in \mathbf{N}$ is the number of default event windows. The event localization submodule ψ_{loc}

is likewise implemented using a 1D convolution layer across the entire data volume.

4.2.1.7 *Data and event sampling*

The proposed network requires an input tensor $x \in \mathbb{R}^{C \times T}$ containing PSG data in the time segment of size T as well as information about the associated events in the segment. Since the total number of segments in a standard PSG without any event data by far outnumbers the number of segments with event data, we implemented a random sampling of non-event and event classes with the sampling probability of class k inversely proportional to the number of classes, such that $p_k = \frac{1}{K+1}$, $k = [0 \mathinner{..} K]$, where $k = 0$ is the default (non-event) class. At training step t, we thus sample a class k and afterwards randomly sample a single class k event ε_k between all class k events. Finally, we extract a segment of PSG data of size $C \times T$ with start of segment in the interval $[\bar{\varepsilon}_k - T, \bar{\varepsilon}_k + T]$, where $\bar{\varepsilon}_k$ is the sample midpoint of ε_k. This ensures that each $\mathbf{x}$ overlaps 50% with at least one associated event.

4.2.1.8 *Optimization of network parameters*

The network parameters were optimized using mini-batch stochastic gradient descent with initial learning rate of 10^{-3} and a momentum of 0.9. Mini-batches were balanced with respect to the detected classes. The optimization of the network was performed with respect to the same loss function described in [113], [114], and the network was trained until convergence determined by no decrease in the loss on the EVAL set over 10 epochs of TRAIN data. We also employed learning rate decay with a factor of 2 every 5 epochs of non-decreasing EVAL loss.

We used a worst negative mining approach with a positive/negative sample ratio of 3.

4.2.1.9 *Experimental setups*

In this study, we examined two different experimental setups:

EXPERIMENT A First, we investigated the differences in predictive performance using a static vs. a dynamic default event window size. This was realized by running six separate training runs with $\tau \in \{3, 5, 10, 15, 20, 30\} \times f_s$, as well as a single training run where f was evaluated for all τ in $\{3, 5, 10, 15, 20, 30\} \times f_s$. The best performing model was determined by evaluating F1 score on the EVAL set for both LM and Ar detection.

EXPERIMENT B Second, we tested a network where we added a recurrent processing block ϕ_R after the feature extraction block ϕ_T as shown in grey in Table 4.2. We considered a single bidirectional gated recurrent unit (bGRU) layer with $\tilde{C}$ units. Predictions were evaluated across multiple time-scales $\tau \in \{3, 5, 10, 15\} \times f_s$.

All experiments were implemented in PyTorch 1.0 [165], [166].

4.2.1.10 *Performance metrics*

All models were evaluated on the EVAL and TEST sets using precision (Pr), recall (Re), and F1 scores (F1):

$$\mathrm{Pr} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad \mathrm{Re} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{4.1}$$

$$\mathrm{F1} = 2\frac{\mathrm{Pr}*\mathrm{Re}}{\mathrm{Pr}+\mathrm{Re}} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}, \tag{4.2}$$

where TP, FP, and FN, are the number of true positives, false positives and false negatives, respectively.

4.2.2 *Results and discussion*

Shown in Figures 4.2a and 4.2b are the F1 scores as a function of IoU and the classification threshold θ_{clf} for both the LM and Ar detection models. It is apparent that both models perform best with a minimum overlap (IoU = 0.1) with their respective annotated events, and do not benefit from increasing the overlap. This might be caused by the annotated events being imprecise and not by issues with the model itself. For example, it is not uncommon to only mark the beginning of an event in standard sleep scoring software, as the duration will automatically be annotated by a default length. . Future studies will be able to confirm this by either collecting a precisely annotated cohort, or by investigating the average start time and duration discrepancies between annotated and predicted events.

3 s *for Ars, and* 0.5 s *for LM are the minimum durations as defined by AASM [43]*

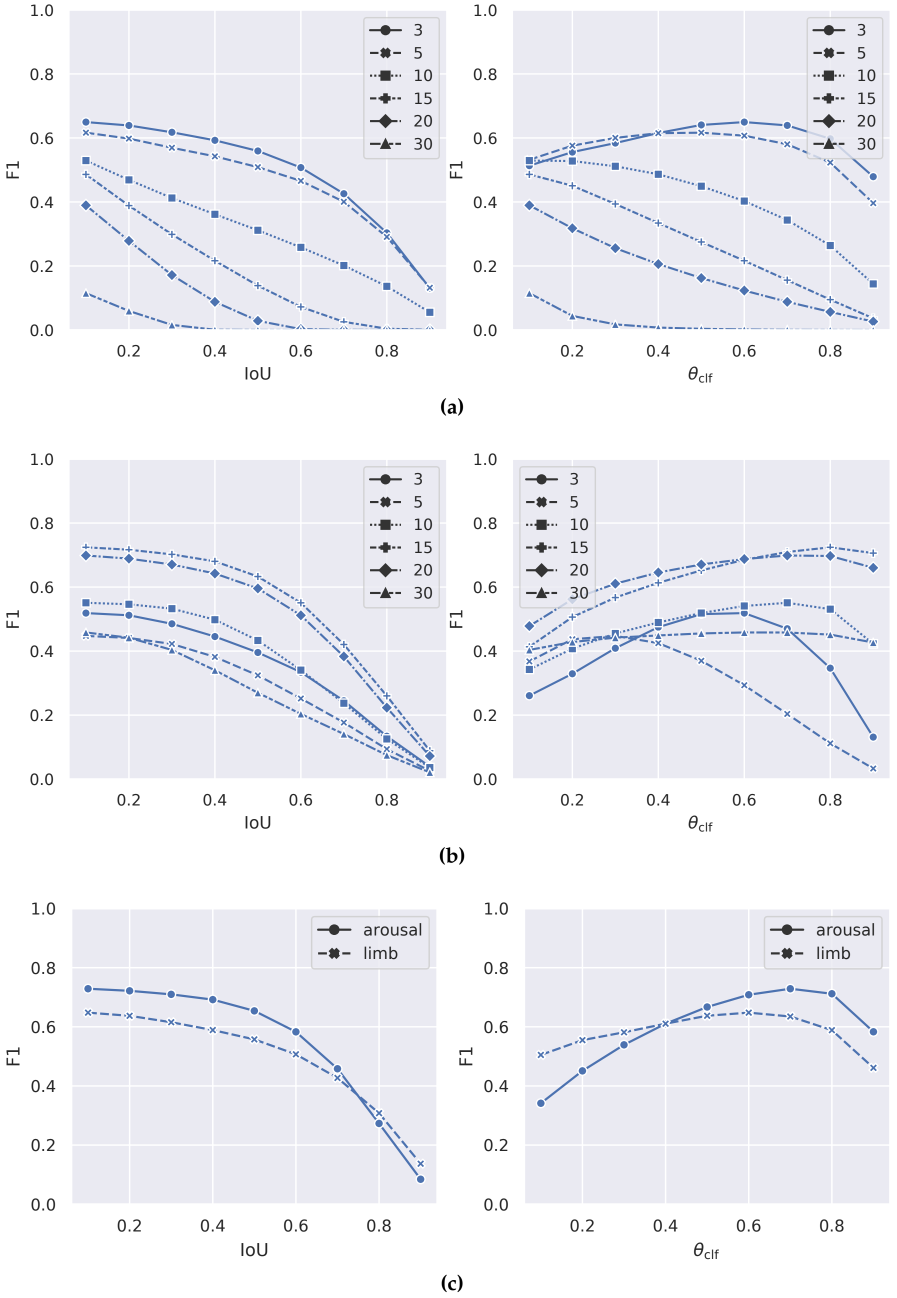
It is also apparent from Figures 4.2a and 4.2b that both detection models benefit from imposing a strict classification threshold. Specifically, LM detection performance as measured by F1 was highest with $\theta_{\mathrm{clf}} = 0.6$, while maximum Ar detection performance was attained with an even higher θ_{clf} of 0.8.

By allowing for multiple time-scales in the dynamic models, shown in Figure 4.2c, we hypothesized that dynamic default event windows would allow for more flexibility and thus better predictive performance. However, we observed no significant differences between the optimal static window and the dynamic window model.

Shown in Figure 4.3 are the performance curves for the RNN (bidirectional GRU) version of the proposed model for each of the two event detection tasks. While the optimal IoU and θ_{clf} points are unchanged from the static/dynamic models presented in Figure 4.2, the optimal F1 value for Ar detection is increased by incorporating temporal dependencies in the model. The reverse is true for LM detection, which saw a slight decrease in predictive performance caused by lower precision. Future work should consider optimizing predictive performance by investigating the effects of varying the number of bGRU layers and the number of hidden units in ϕ_{R}, since this was not performed here.

See Table 4.3

Application of the optimal models on the TEST data is shown in Table 4.3. With the given architecture of f and the given labels and input data in TRAIN, LM detection was maximal for the model with a static/dynamic window, while adding a recurrent module only positively impacted Ar prediction. Precision and recall decreased for LM detection when adding ϕ_{R}, while precision increased and recall decreased for Ar detection. An example visualization of the joint distribution of F1 scores obtained from the dynamic model applied to the TEST data is shown in Figure 4.4. While some outliers



(a)

(b)

(c)

Figure 4.2: Experiment A: Optimizing IoU and θ_{clf} in static models on the EVAL set by varying default event window size in seconds in $\{3, 5, 10, 15, 20, 30\}$ **(a)-(b)**. Left panels show the IoU vs. F1 score, while right panels show classification threshold θ_{clf} against F1 score. **(a)** LM model. Here, the model performs best for IoU $= 0.1$ and $\theta_{\text{clf}} = 0.6$ using a window size of $\tau = 3\,\text{s} \times f_s$. **(b)** Ar model. Here, the model performs best for IoU $= 0.1$ and $\theta_{\text{clf}} = 0.8$ using a window size of $\tau = 15\,\text{s} \times f_s$. **(c)** Dynamic models show optimal performance for IoU $= 0.1$ and $\theta_{\text{clf}} = 0.7$ and $\theta_{\text{clf}} = 0.6$ for Ar and LM detection, respectively.

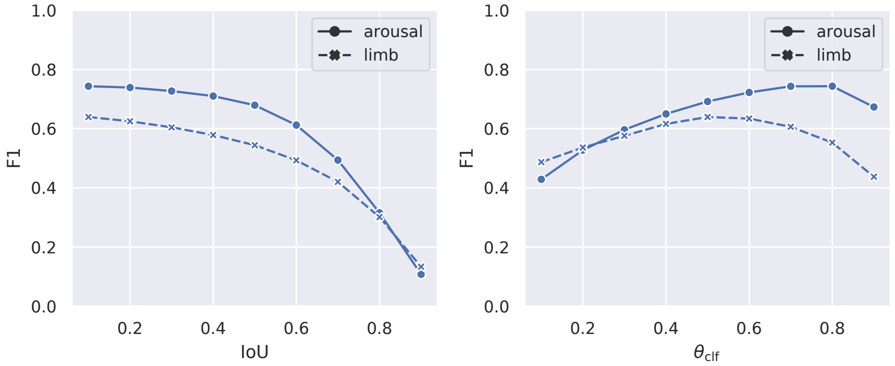


Figure 4.3: Experiment B. F1 performance on the EVAL set as a function of IoU and θ_{clf} for Ar and LM detection when adding the ϕ_R module. Best performance is seen for IoU $= 0.1$ for both Ar and LM detection, and $\theta_{clf} = 0.6$ and $\theta_{clf} = 0.8$ for LM and Ar detection, respectively.

are readily observable, especially for LM detection, the majority of subject F1 scores follows an approximate bivariate normal distribution.

Subset partitions were reasonably well-distributed with no significant differences between key variables, see Table 4.1. An exception is the AHI, although the associated effect is small and most likely a result of the low sample size in EVAL compared to TRAIN and TEST. It is noted, that although AHI, arousal index (ArI), and PLMI are not normally distributed and summarizing these variables with standard deviations is invalid, it is nevertheless standard practice in sleep medicine and thus presented the same way here. We performed little data cleaning in order to provide as much data and variation to the deep learning model as possible, however, future efforts should explore and apply inclusion criteria such as minimal total sleep time, artifact detection and removal of studies with severe artifacts. We did impose a trivial lower bound on the number of scored events (>0) for a PSG to be included in this study, but stricter requirements could potentially improve model performance.

In this work, we investigated somatic PSG events present in multiple signal modalities instead of EEG-specific events, which required changes to the network architecture. Specifically, we kept the signal modality encoded in the first dimension of the tensor propagated through the network, which allowed for the use of 1D convolutional operators. By performing 1D convolutions and keeping the channel information in the feature maps instead of keeping them as separate dimensions and performing 2D convolutions as proposed in [113], [114], we simplify and reduce the number of computations and training time by a factor $\propto$ C.

However, we did not investigate the effects of modeling the conditional probability of Ar and LM occurrence, but the proposed architecture is versatile enough to detect both events jointly as well as separately. Previous work also suggest that detecting multiple objects at the same time is of high interest and leads to (at least) non-inferior performances [113], [114], [160]–[162].

Additionally, we speculated that the temporal dynamics of the PSG signals were important for optimal event detection performance. Although the effects were small, the F1 score in Ar detection increased when adding an RNN module to the network before the detection module. However, this was not

Table 4.3: Application of optimized models on TEST data.

Model	F1	Pr	Re
LM, static	0.648 ± 0.148	0.631 ± 0.181	0.720 ± 0.141
Ar, static	0.727 ± 0.102	0.706 ± 0.113	0.771 ± 0.132
LM, dynamic	0.647 ± 0.148	0.627 ± 0.181	0.722 ± 0.140
Ar, dynamic.	0.729 ± 0.102	0.699 ± 0.115	0.785 ± 0.131
LM, RNN	0.639 ± 0.147	0.606 ± 0.180	0.727 ± 0.126
Ar, RNN	0.749 ± 0.105	0.772 ± 0.107	0.748 ± 0.138

Data are shown as subject-averaged F1, precision (Pr) and recall (Re) with associated standard deviations. Top four rows correspond to Experiment A, while bottom two rows correspond to Experiment B. Ar: arousal; LM: leg movement; RNN: recurrent neural network.

the case for LM detection, which is most likely due to the different temporal and physiological characteristics of the two events in question.

Future efforts will address the fact that events are mutually exclusive in the current modeling scheme, given a certain default event window size. However, it is common to see Ars and LMs as a result of one another, and thus, if the window size is too small, a more unlikely event, as measured by classification threshold and IoU, will be removed even if it matches up to a specific true event of a certain class.

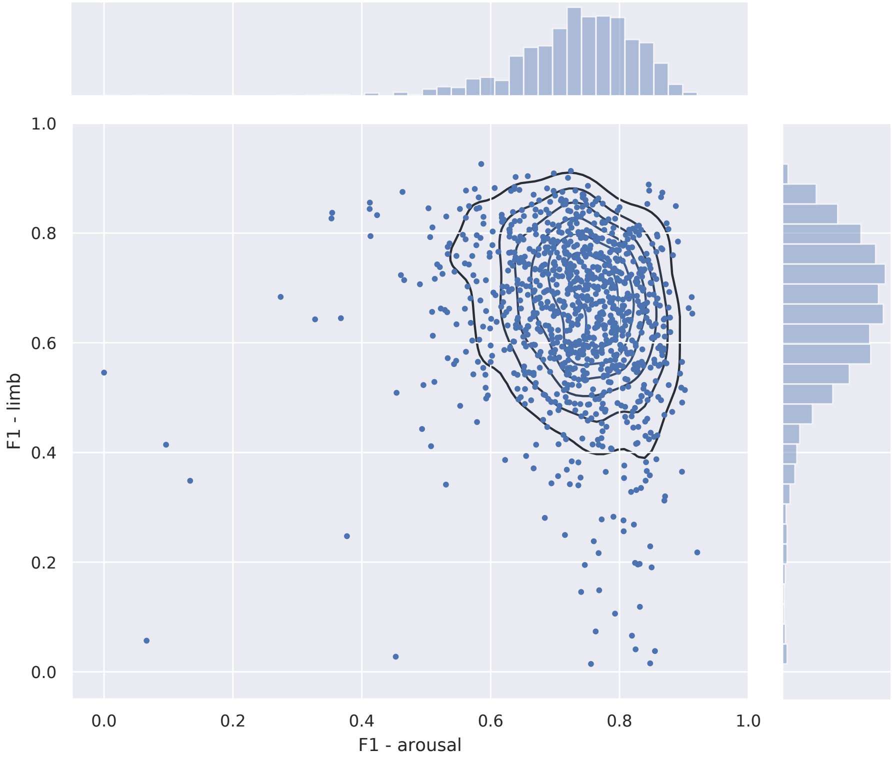


Figure 4.4: Visualization of F1 scores for both Ar and LM detection using the dynamic model.

4.3 PAPER VI: A MULTI-MODAL SLEEP EVENT DETECTION MODEL FOR CLINICAL SLEEP ANALYSIS

STUDY OBJECTIVE: Clinical sleep analysis require manual analysis of sleep patterns for correct diagnosis of sleep disorders. Several studies show significant variability in scoring discrete sleep events. We wished to investigate, whether an automatic method could be used for detection of arousals (Ar), leg movements (LM) and sleep disordered breathing (SDB) events, and if the joint detection of these events performed better than having three separate models.
METHODS: We designed a single deep neural network architecture to jointly detect sleep events in a polysomnogram. We trained the model on 1653 recordings of individuals, and tested the optimized model on 1000 separate recordings. The performance of the model was quantified by F1, precision, and recall scores, and by correlating index values to clinical values using Pearson's correlation coefficient.
RESULTS: F1 scores for the optimized model was 0.70, 0.63, and 0.62 for Ar, LM, and SDB, respectively. The performance was higher, when detecting events jointly compared to corresponding single-event models. Index values computed from detected events correlated well with manual annotations ($r^2 = 0.73$, $r^2 = 0.77$, $r^2 = 0.78$, respectively).
CONCLUSION: Detecting arousals, leg movements and sleep disordered breathing events jointly is possible, and the computed index values correlates well with human annotations.

Clinical sleep analysis is currently performed manually by experts based on guidelines from the AASM detailed in the AASM Scoring Manual [43]. The guidelines detail both technical and clinical best practices for setting up and recording PSGs, which are overnight recordings of various electrophysiological signals, such as EEG, EOG, chin and leg EMG, ECG, respiratory inductance plethysmography from the thorax and abdomen, oronasal pressure, and blood oxygen levels.

Based on these signals, expert technicians analyse and score the PSG for sleep stages [W, REM sleep, N1, N2, and N3], and sleep micro-events summarized in key metrics, such as the AHI (number of apneas and hypopneas per hour of sleep), the PLMI (number of period leg movements per hour of sleep), and the ArI (number of arousals per hour of sleep).

Arousals are defined as abrupt shifts in EEG frequencies towards alpha, theta, and beta rhythms for at least 3 s with a preceding period of stable sleep of at least 10 s. During REM sleep, where the background EEG shows similar rhythms, arousal scoring requires a concurrent increase in chin EMG lasting at least 1 s. LMs should be scored in the leg EMG channels, when there is an increase in amplitude of at least 8 µV above baseline level with a duration between 0.5 s to 10 s. A PLM series is then defined as a sequence of 4 LMs, where the time between LM onsets is between 5 min to 90 min. Apneas are generally scored when there is a complete (⩾90 % of pre-event baseline) cessation of breathing activity either due to a physical obstruction (obstructive apnea) or due to an underlying disruption in the central nervous system control (central apnea) for at least 10 s. When the breathing is only partially reduced (⩾30 % of pre-event baseline) and the duration of the excursion is ⩾10 s, the event is scored as a hypopnea if there is either a ⩾4 % oxygen desaturation or a ⩾3 % oxygen desaturation coupled with an Ar.

However, several studies have shown significant variability in the scoring of both sleep stages [23], [24], [26], [28], [29], [67], [167] and sleep micro-events [21], [22], [25], [27], [68]–[71]. This has prompted extensive research into automatic methods for classifying sleep stages in large-scale studies [72], [73], [83], [85], [87]–[89], [99], while the research in automatic arousal [115], [158], [168] and LM [159] detection on a similar scale is limited. Biswal *et al.* recently proposed a model based on a combination of recurrent and convolutional neural networks, where the same architecture was used for sleep stage classification, AHI and limb movement index (LMI) prediction. They trained their model using 9000 PSG recordings and evaluated the performance on the three tasks on a held out test set containing 1000 PSGs. However, this model was trained in separate runs for each downstream task; furthermore, post-processing was performed on the event predictions (apneas, limb movements).

In this study, we introduce the MSED model for joint detection of sleep micro-events, in this case Ars, SDB, and LMs. The model is based on recent advances in machine learning and challenges current state of the art methods by directly classifying and localizing sleep micro-events in the PSG signals at the same time.

4.3.1 *Data*

We collected PSGs from the MrOS Sleep Study, an ancillary part of the larger Osteoporotic Fractures in Men Study. The main goal of the study is to research and discover connections between sleep disorders, skeletal fractures, and cardiovascular disease and mortality in community-dwelling older (>65 years) [103]–[105]. Of the original 5994 study participants, 3135 subjects were enrolled at one of six sites in the USA for a comprehensive sleep assessment, while 2909 of these underwent a full-night in-home PSG recording, The PSG studies were subsequently scored by certified sleep technicians. Sleep stages were scored into stages 1, 2, 3, 4 and REM, while stages 3 and 4 combined into slow wave sleep (SWS) according to R&K rules [106]. Ars were scored as abrupt increases in EEG frequencies lasting at least 3 s according to ASDA rules [169]. Apneas were defined as complete or near complete cessation of airflow lasting more than 10 s with an associated 3 % or greater SaO_2 desaturation, while hypopneas were based on a clear reduction in breathing of more than 30 % deviation from baseline breathing lasting more than 10 s, and likewise assocated with a greater than 3 % SaO_2 desaturation. While the scoring criteria for scoring LMs are not explicitly available for the MrOS Sleep Study, the prevailing standard at the time of the study was to score LMs following an increase in leg EMG amplitude of more than 8 µV above resting baseline levels for at least 0.5 s, but shorter than 10 s [170].

4.3.1.1 *Subset demographics and partitioning*

We used a total of 2853 PSG studies downloaded from the NSRR [107], [108], which we partitioned into a training set ($\mathcal{D}_{\text{TRAIN}}$, $n_{\text{train}} = 1653$), a validation set ($\mathcal{D}_{\text{EVAL}}$, $n_{\text{eval}} = 200$), and a final testing set ($\mathcal{D}_{\text{TEST}}$, $n_{\text{test}} = 1000$). Key demographics and PSG-related variables for each subset are shown as mean ± standard deviation with range in parenthesis in Table 4.4.

Table 4.4: MrOS demographics by subset.

	$\mathcal{D}_{\text{TRAIN}}$	$\mathcal{D}_{\text{EVAL}}$	$\mathcal{D}_{\text{TEST}}$	p-value
n	1653	200	1000	-
Age, years	76.4 ± 5.6 [67.0 – 90.0]	76.8 ± 5.4 [68.0 – 90.0]	76.4 ± 5.3 [67.0 – 90.0]	0.404
BMI, $\mathrm{kg\,s^{-2}}$	27.3 ± 3.9 [16.0 – 47.0]	27.0 ± 3.6 [19.0 – 40.0]	27.0 ± 3.7 [17.0 – 45.0]	0.247
TST, min	357.3 ± 69.0 [54.0 – 615.0]	354.0 ± 69.1 [108.0 – 503.0]	353.6 ± 68.7 [62.0 – 572.0]	0.312
SL, min	22.9 ± 25.6 [1.0 – 349.0]	21.6 ± 23.0 [1.0 – 135.0]	25.1 ± 32.1 [1.0 – 402.0]	0.284
REML, min	109.5 ± 77.9 [0.0 – 578.0]	103.5 ± 70.0 [10.0 – 413.0]	107.2 ± 75.3 [3.0 – 590.0]	0.466
WASO, min	116.7 ± 67.1 [11.0 – 462.0]	119.0 ± 70.8 [15.0 – 372.0]	112.9 ± 65.0 [6.0 – 458.0]	0.471
SE, %	75.9 ± 12.1 [17.0 – 97.0]	75.5 ± 12.3 [37.0 – 96.0]	76.4 ± 11.8 [26.0 – 98.0]	0.690
N1, %	6.8 ± 4.1 [0.0 – 31.0]	7.0 ± 4.5 [0.0 – 28.0]	6.9 ± 4.7 [1.0 – 58.0]	0.968
N2, %	62.7 ± 9.5 [28.0 – 89.0]	62.0 ± 9.7 [30.0 – 90.0]	62.8 ± 10.0 [21.0 – 95.0]	0.451
N3, %	11.4 ± 9.0 [0.0 – 55.0]	11.8 ± 9.7 [0.0 – 55.0]	11.1 ± 9.0 [0.0 – 57.0]	0.638
REM, %	19.2 ± 6.5 [0.0 – 44.0]	19.4 ± 7.2 [0.0 – 41.0]	19.3 ± 6.7 [0.0 – 42.0]	0.894
ArI, $\mathrm{h^{-1}}$	23.5 ± 11.8 [3.0 – 87.0]	23.4 ± 11.0 [4.0 – 77.0]	23.8 ± 11.8 [4.0 – 102.0]	0.661
AHI, $\mathrm{h^{-1}}$	13.5 ± 13.9 [0.0 – 83.0]	13.6 ± 13.3 [0.0 – 59.0]	14.2 ± 15.5 [0.0 – 89.0]	0.907
PLMI, $\mathrm{h^{-1}}$	35.4 ± 37.1 [0.0 – 233.0]	36.6 ± 39.0 [0.0 – 178.0]	36.0 ± 37.7 [0.0 – 175.0]	0.993

Data are shown as means ± standard deviations (range) across PSGs. Variables were tested with Kruskal-Wallis H-tests. Significant p-values at significance level $\alpha = 0.05$ are highlighted in bold. BMI: body-mass index; TST: total sleep time; SL, sleep latency, REML: REM sleep latency; WASO: wake after sleep onset; SE, sleep efficiency, N1: non-rapid eye movement stage 1; N2: non-rapid eye movement stage 2; N3: non-rapid eye movement stage 3; REM: rapid eye movement; ArI: arousal index; AHI: apnea-hypopnea index; PLMI: periodic leg movement index.

4.3.1.2 *Signal and events*

For this study, we considered three PSG events: Ars, LMs, and SDB events, which includes all forms of apneas (obstructive and central) and hypopneas. These event types are each based on a specific set of electrophysiological channels from the PSG, and as such, we extracted left and right central EEG (C3 and C4), left and right EOG, left and right chin EMG, left and right leg EMG, nasal pressure, and respiratory inductance plethysmography from the thorax and abdomen. EEG and EOG channels were referenced to the contralateral mastoid process, while a chin EMG was synthesized by subtracting the right chin EMG from the left chin EMG.

Apart from the raw signal data, we also extracted onset time relative to the study start time and duration times for each event type in each PSG.

4.3.2 *Methods*

NOTATION We denote by $\llbracket a, b \rrbracket$ the set of integers $\{n \in \mathbb{N} \mid a \leqslant n \leqslant b\}$ with $\llbracket N \rrbracket$ being shorthand for $\llbracket 1, N \rrbracket$, and by $n \in \llbracket N \rrbracket$ the nth sample in $\llbracket N \rrbracket$. A segment of PSG data is denoted by $\mathbf{x} \in \mathbb{R}^{C \times T}$, where C, T is the number of channels and the duration of the segment in samples, respectively.

The corresponding set of N_t true events for the segment is denoted by $\varepsilon^t = \left\{\left(\rho_i^t, \delta_i^t\right) \in \mathbb{R}_+^2 \mid i \in [\![N_t]\!]\right\}$, where ρ, δ are the center point and duration, respectively, of the *i*th event. By **ffl** $\in \mathcal{D}_*$ we denote a sample in either one of the three subsets. In the description of the network architecture, we have omitted the batch dimension from all calculations for brevity.

4.3.2.1 *Model overview*

Given an input set $\chi = \{\mathbf{x}, \varepsilon^t\} \in \mathbb{R}^{C \times T} \times \mathbb{R}_+^{N_t \times 2}$ containing PSG data with C channels and T time steps, and true events ε, the goal of the model is to detect any possible events in the segment, where, in this context, detection covers both classification *and* localization of any event in the segment space.

To accomplish this, the model generates a set of *default event windows* $\varepsilon^d = \left\{\left(\rho_j^d, \delta_j^d\right) \in \mathbb{R}_+^2 \mid j \in [\![N_d]\!]\right\}$ for the current segment, and match each true event to a default event window if their intersection-over-union (IoU) is at least 0.5.

At test time, we generate predictions over the default event windows and use a non-maximum suppression procedure to select between the candidate predictions. For a given class *k*, the procedure is as follows. First, the predictions are sorted according to probability of the event, which is above a threshold θ_k. Then, using the most probable prediction as an anchor, we sequentially evaluate the IoU between the anchor and the remaining candidate predictions, removing those with IoU >= 0.5.

The output of the model is thus the set $\{\mathbf{p}, \mathbf{y}\}$ containing the predicted class probabilities along with the corresponding onsets and durations.

4.3.2.2 *Signal processing pipeline*

We resampled all signals to a common sampling frequency of $f_s = 128\,\text{Hz}$ using a poly-phase filtering approach (Kaiser window, $\beta = 5.0$). Based on recommended filter specifications from the AASM, we designed Butterworth IIR filters for four sets of signals. EEG and EOG channels were filtered with a 2nd order filter with a 0.3 Hz–35 Hz passband, while chin and leg EMG channels were filtered with a 4th order high-pass filter with a 10 Hz cut-off frequency. The nasal pressure channel was filtered with a 4th order high-pass filter with a 0.03 Hz cut-off frequency, while the thoracoabdominal channels were filtered with a 2nd order with a 0.1 Hz–15 Hz passband.

All filters were implemented using the zero-phase method, which sequentially applies the filter in the forward direction, and then in the backwards direction. This accounts for the non-linear phase response and subsequent frequency-dependent group delay inherent in IIR filters, but also effectively squares the magnitude response of the filter.

Filtered signals were subsequently standardized by

$$\mathbf{x}^{(i)} = \frac{\hat{\mathbf{x}}^{(i)} - \boldsymbol{\mu}^{(i)}}{\boldsymbol{\sigma}^{(i)}}, \tag{4.3}$$

where $\hat{\mathbf{x}}^{(i)} \in \mathbb{R}^{C \times T}$ is the raw matrix containing C input channels and T samples, and $\boldsymbol{\mu}^{(i)}, \boldsymbol{\sigma}^{(i)} \in \mathbb{R}^C$ are the mean and standard deviation vectors for the *i*th PSG, respectively. This is a common approach in computer vision tasks, and beneficial to ensure a proper gradient propagation through a deep neural network [171].

4.3.2.3 *Target encoding*

For each data segment, target event classes $\boldsymbol{\pi} \in \mathbb{R}^{N_m \times K}$ generated by one-hot encoding, while the target detection variable containing the onset and duration times $\mathbf{t} \in \mathbb{R}^{N_m \times 2}$ was encoded as

$$t_i = \left(\frac{\rho_i^m - \rho_j^d}{\delta_j^d}, \log \frac{\delta_i^m}{\delta_j^d} \right), \quad i \in [\![N_m]\!], \, j \in [\![N_d]\!], \tag{4.4}$$

where ρ_i^m is the center point of the true event matched to a default event window ρ_j^d, and δ_i^m and δ_j^d are the corresponding durations of the true and default events.

4.3.2.4 *Data sampling*

As the total number of default event windows in a data segment N_d most likely will be much higher than the number of event windows matched to a true event, i.e. $N_d \gg N_m$, we implemented a similar random data sampling strategy as in [115]. At training step t, a given PSG record r has a certain number of associated number of Ar, LM, and SDB events ($n_{\text{Ar}}, n_{\text{LM}}, n_{\text{SDB}}$, respectively). We randomly sample a class k with equal probability $p_k = \frac{1}{K}$, whilst disregarding the negative class, since this class is most likely over-represented in the data segment. Given the class k, we randomly sample an event ε_k with probability $p(\varepsilon_k) \propto n_k$, and afterwards, we extract a segment of data of size $C \times T$, where the start of the segment is sampled from $[\bar{\varepsilon}_k - T, \bar{\varepsilon}_k + T]$, where $\bar{\varepsilon}_k$ is the sample midpoint of the event ε_k, thereby ensuring at least 50 % overlap with at least one event associated with the data segment.

We found that this approach to sampling data segments with a large ratio of negative to positive samples to be beneficial in all our experiments, when monitoring the loss on the validation set.

4.3.2.5 *Network architecture*

Similar to the architecture described in [158], we designed a splitstream network architecture for the differentiable function Φ, where each stream is responsible for the bulk feature extraction for a specific event class. For the given problem of detecting Ars, LMs, and SDBs, the network contains three streams: the Ar stream takes as input the EEGs, the EOGs, and the chin EOG signals for a total of $C_{\text{Ar}} = 5$ channels; the LM stream receives the $C_{\text{LM}} = 2$ leg EMG signals; and the SDB stream receives the nasal pressure and the thoracoabdominal signals for a total of $C_{\text{SDB}} = 3$ channels. An overview of the network architecture is shown graphically in Figure 4.5.

4.3.2.6 *Stream specifics*

Each stream is comprised of two components. First, a mixing module $\varphi_{\text{mix}} : \mathbb{R}^{C_* \times T} \to \mathbb{R}^{C_* \times T}$ computes a non-linear mixing of the C channels using a set of C single-strided 1-dimensional filters $\mathbf{w} \in \mathbb{R}^{C \times C}$ and ReLU activation [172], such that $\varphi_{\text{mix}}(\mathbf{x}) = \max\{0, \mathbf{w} \otimes \mathbf{x} + \mathbf{b}\}$, where the max operation introduces the non-linearity, $\otimes$ is the conv operator over the C feature maps, and $\mathbf{b} \in \mathbb{R}^C$ is a bias vector (in this case $\mathbf{b} = 0$). Second, the output activations from φ_{mix} are used as input to a deep neural network module $\varphi_{\text{feat}} : \mathbb{R}^{C_* \times T} \to \mathbb{R}^{f' \times T'}$, which transforms the input feature maps to a $f' \times T'$ feature space with a temporal dimension reduced by a factor of $\frac{T}{T'}$. The

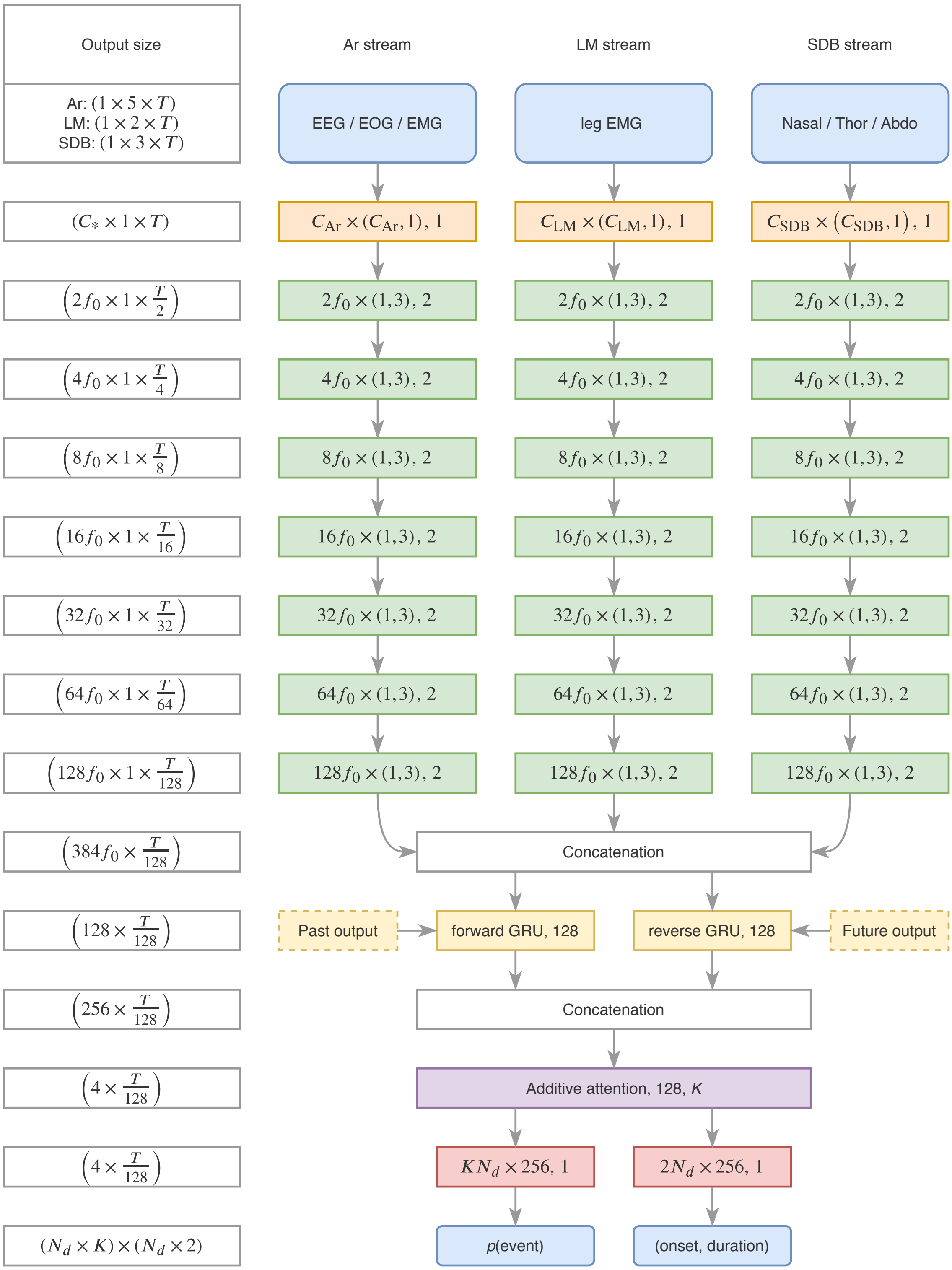

Output size
Ar stream
LM stream
SDB stream
Ar: $(1\times5\times T)$
LM: $(1\times2\times T)$
SDB: $(1\times3\times T)$
EEG / EOG / EMG
leg EMG
Nasal / Thor / Abdo
$(C_*\times1\times T)$
$C_{\mathrm{Ar}}\times(C_{\mathrm{Ar}},1)$, 1
$C_{\mathrm{LM}}\times(C_{\mathrm{LM}},1)$, 1
$C_{\mathrm{SDB}}\times(C_{\mathrm{SDB}},1)$, 1
$\left(2f_0\times1\times\frac{T}{2}\right)$
$2f_0\times(1,3)$, 2
$\left(4f_0\times1\times\frac{T}{4}\right)$
$4f_0\times(1,3)$, 2
$\left(8f_0\times1\times\frac{T}{8}\right)$
$8f_0\times(1,3)$, 2
$\left(16f_0\times1\times\frac{T}{16}\right)$
$16f_0\times(1,3)$, 2
$\left(32f_0\times1\times\frac{T}{32}\right)$
$32f_0\times(1,3)$, 2
$\left(64f_0\times1\times\frac{T}{64}\right)$
$64f_0\times(1,3)$, 2
$\left(128f_0\times1\times\frac{T}{128}\right)$
$128f_0\times(1,3)$, 2
$\left(384f_0\times\frac{T}{128}\right)$
Concatenation
$\left(128\times\frac{T}{128}\right)$
Past output
forward GRU, 128
reverse GRU, 128
Future output
$\left(256\times\frac{T}{128}\right)$
Concatenation
$\left(4\times\frac{T}{128}\right)$
Additive attention, 128, K
$\left(4\times\frac{T}{128}\right)$
$KN_d\times256$, 1
$2N_d\times256$, 1
$(N_d\times K)\times(N_d\times2)$
p(event)
(onset, duration)

Figure 4.5: MSED network architecture. The left column shows the output dimensions for each operation as (number of filters[x singleton] x time steps). Each stream on the right (green) processes a separate set of input channels (blue, top), the results of which are concatenated before the bGRU (yellow). The outputs from the additive attention layer (purple) are convolved in the final classification and localization layers (red) to output the probabilities for each event class, and the predicted onset and duration of each event (blue, bottom). Convolution layers (orange, green, red) are detailed as [number of feature maps x kernel size, stride]. Recurrent layer (yellow) shows the direction and number of hidden units. Additive attention layer (purple) is described with the number of hidden and output units.

feature extraction module φ_{feat} is realized using $k_{\max}$ successive conv operations with an increasing number of filters $f' = f_0 2^{k-1}$, $k \in [\![k_{\max}]\!]$, where f_0 is a tunable base filter number. Each conv feature map is normalized using batch normalization (BN) [116], such that if $\tilde{\mathbf{z}} \in \mathbb{R}^{f' \times T'}$ denotes the output from a conv operation, the subsequent normalized version is computed as

$$\mathbf{z} = \gamma \frac{\tilde{\mathbf{z}} - \mathrm{E}[\tilde{\mathbf{z}}]}{\sqrt{\mathrm{Var}[\tilde{\mathbf{z}}] + \epsilon}} + \beta, \tag{4.5}$$

where $\mathrm{E}[\tilde{\mathbf{z}}] \in \mathbb{R}^{f'}$, $\mathrm{Var}[\tilde{\mathbf{z}}] \in \mathbb{R}_+^{f'}$ is the expectation and variance over the temporal dimension of each feature map, ϵ is a small constant, and $\{\gamma, \beta\} \in \mathbb{R}^{f'} \times \mathbb{R}^{f'}$ are learnable parameters representing the mean and bias for each feature map. Each normalized conv output is subsequently activated using ReLU.

4.3.2.7 *Feature fusion for sequential processing*

The outputs from the three feature extraction streams are subsequently fused by concenating each output vector $\mathbf{z}_*$ into a combined feature vector $\mathbf{z} = (\mathbf{z}_{\text{ar}}, \mathbf{z}_{\text{lm}}, \mathbf{z}_{\text{sdb}}) \in \mathbb{R}^{3f' \times T'}$. We introduce sequential modeling of the feature vectors using a bGRU [117], which has the advantage over other RNN-based models such as the LSTM of having fewer trainable parameters while still being powerful enough to model complex, temporal relationships [173]. The output of the GRU for timestep *t* is a vector $\mathbf{h}_t = \left(\mathbf{h}_t^f, \mathbf{h}_t^b\right) \in \mathbb{R}^{2n_h}$ containing the concatenated outputs from the forward (f) and backward (b) directions. Each directional feature vector is calculated as a weighted combination of a gated new input $\mathbf{n}_t$ and the feature vector from the previous timestep $\mathbf{h}_{t-1}$

$$\mathbf{h}_t^* = (1 - \mathbf{u}_t) \otimes \mathbf{n}_t + \mathbf{u}_t \otimes \mathbf{h}_{t-1}. \tag{4.6}$$

The update gate $\mathbf{u}_t$ and gated new input $\mathbf{n}_t$ are computed as

$$\mathbf{u}_t = \sigma\left(\mathbf{W}_u^z \mathbf{z}_t + \mathbf{b}_u^z + \mathbf{W}_u^h \mathbf{h}_{t-1} + \mathbf{b}_u^h\right), \tag{4.7}$$

$$\mathbf{n}_t = \tanh\left(\mathbf{W}_n^z \mathbf{z}_t + \mathbf{b}_n^z + \mathbf{r}_t \otimes \left(\mathbf{W}_n^h \mathbf{h}_{t-1} + \mathbf{b}_n^h\right)\right), \tag{4.8}$$

where $\mathbf{W}_*^*, \mathbf{b}_*^*$ are weight matrices and bias vectors, respectively, and $\mathbf{r}_t$ is a reset gate computed as

$$\mathbf{r}_t = \sigma(\mathbf{W}_r^z \mathbf{z}_t + \mathbf{b}_r^z + \mathbf{W}_h^r \mathbf{h}_{t-1} + \mathbf{b}_h^r). \tag{4.9}$$

4.3.2.8 *Additive attention*

The attention mechanism is a powerful technique to introduce a way for the network to focus on relevant regions and disregard irrelevant regions of a data sample, and is a key part of the highly successful Transformer model [174] and the subsequent state-of-the-art BERT model for natural language processing [175]. In this work, we implemented a simple, but powerful, *additive attention* mechanism [176], which computes *context*-vectors $\mathbf{c} \in \mathbb{R}^{2n_h}$ for each event class as the weighted sum of the feature vector outputs $\mathbf{h} \in \mathbb{R}^{2n_h \times T'}$ from the φ_h. Formally, attention is computed as

$$\mathbf{c} = \mathbf{h} \cdot \boldsymbol{\alpha} = \sum_{t=1}^{T'} \mathbf{h}_t \alpha_t, \tag{4.10}$$

where T' is the reduced temporal dimension, $\mathbf{h}_t$ is the feature vector for time step t, and $\boldsymbol{\alpha}_t \in \mathbb{R}^K$ is the attention weight computed as

$$\boldsymbol{\alpha}_t = \frac{\exp(\tanh(\mathbf{h}_t \mathbf{W}_u)\mathbf{W}_a)}{\sum_\tau^{T'} \exp(\tanh(\mathbf{h}_\tau \mathbf{W}_u)\mathbf{W}_a)}. \tag{4.11}$$

Here, $\mathbf{W}_u \in \mathbb{R}^{2n_h \times n_a}$ and $\mathbf{W}_a \in \mathbb{R}^{n_a \times K}$ are linear mappings of the feature vectors, and tanh is the hyperbolic tangent function.

4.3.2.9 *Detection*

The final event classification and localization is handled by two modules, $\psi_{\text{clf}} : \mathbb{R}^{2n_h \times K} \to \mathbb{R}^{N_d \times K}$ and $\psi_{\text{loc}} : \mathbb{R}^{2n_h \times K} \to \mathbb{R}^{N_d \times 2}$, respectively. The classification module $\psi_{\text{clf}} : \mathbf{c} \mapsto \mathbf{p}$ outputs a tensor $\mathbf{p} \in [0,1]_+^{N_d \times K}$ containing predicted event class probabilities for each default event window. The localization module $\psi_{\text{loc}} : \mathbf{c} \mapsto \mathbf{y}$ outputs a tensor $\mathbf{y} \in \mathbb{R}^{N_d \times 2}$ containing encoded relative onsets and durations for a detected event for each default event window.

4.3.2.10 *Loss function*

Similar to [156], we optimized the network parameters according to a three-component loss function consisting of: i) a localization loss ℓ_{loc}; ii) a positive classification loss ℓ_+, and iii) a negative classification loss ℓ_-, such that the total loss ℓ was defined by

$$\ell = \ell_{\text{loc}} + \ell_+ + \ell_-. \tag{4.12}$$

The localization loss ℓ_{loc} was calculated using a Huber function

$$\ell_{\text{loc}} = \frac{1}{N_+} \sum_{i \in \pi_+} f_H^{(i)} \tag{4.13}$$

$$\mathbf{f}_H = \begin{cases} 0.5(\mathbf{y} - \mathbf{t})^2, & \text{if } |\mathbf{y} - \mathbf{t}| < 1, \\ |\mathbf{y} - \mathbf{t}| - 0.5, & \text{otherwise,} \end{cases} \tag{4.14}$$

where $i \in \pi_+$ yields indices of event windows with positive targets, i. e. event windows matched to an arousal, LM or SDB target, and N_+ is the number of positive targets in the given data segment.

The positive classification loss component ℓ_+ was calculated using a simple cross-entropy over the event windows matched to an arousal, LM, or SDB event:

$$\ell_+ = \frac{1}{N_+} \sum_{i \in \pi_+} \sum_{k \in [\![K]\!]} \pi_k^{(i)} \log p_k^{(i)}, \quad \text{where} \quad p_k^{(i)} = \frac{\exp s_k^{(i)}}{\sum_j \exp s_j^{(i)}}, \tag{4.15}$$

and $\pi_k^{(i)}$, $p_k^{(i)}$, and $s_k^{(i)}$ are the true class probability, predicted class probability, and logit score for the *i*th event window containing a positive sample.

Similar to [113], [114], the negative classification loss ℓ_- was calculated using a hard negative mining approach to balance the number of positive and negative samples in a data segment after matching default event windows to true events [162]. Specifically, this is accomplished by calculating the probability for the negative class (no event) for each unmatched default event window, and then calculating the cross entropy loss using the Z most probable samples. In our experiments, we set the ratio of positive to negative samples as 1:3, such that the calculation of ℓ involves $Z = 3$ times as many negative as positive samples.

We also explored a focal loss objective function for computing ℓ_+ and ℓ_- [177], however, we found that this approach severely deteriorated the ability of the network to accurately detect LM and SDB events compared to using worst negative mining.

4.3.2.11 *Optimization*

The network parameters were optimized using adaptive moment estimation (Adam) according to the loss function described in Equation (4.12) [97]. This algorithms uses first (m) and second (v) moment estimations of gradients to update the model parameters θ of a differentiable function f at time t:

$$m^{(t)} = \beta_1 m^{(t-1)} + (1-\beta_1)\nabla_\theta f^{(t)}\left(\theta^{(t-1)}\right) \tag{4.16}$$

$$v^{(t)} = \beta_2 v^{(t-1)} + (1-\beta_2)\nabla_\theta^2 f^{(t)}\left(\theta^{(t-1)}\right), \tag{4.17}$$

where β_1, β_2 are exponential decay rates for the first and second moment, respectively, ∇ is the gradient vector with respect to θ, and ∇_θ^2 is the Hadamard product $\nabla_\theta f \odot \nabla_\theta f$. The moment vectors are initialized with 0's, which induce a bias towards zero. This can be offset by computing a bias-corrected estimate of each moment vector as

$$\hat{m}^{(t)} = \frac{m^{(t)}}{1-\beta_1^t} \tag{4.18}$$

$$\hat{v}^{(t)} = \frac{v^{(t)}}{1-\beta_2^t}, \tag{4.19}$$

which yields the final update to θ as

$$\theta^{(t)} = \theta^{(t-1)} - \eta\frac{\hat{m}^{(t)}}{\sqrt{\hat{v}^{(t)}} + \epsilon}, \tag{4.20}$$

where η is the learning rate.

4.3.2.12 *Experimental setups*

In our experiments, we fixed the exponential decay rates at $(\beta_1, \beta_2) = (0.9, 0.999)$, the learning rate at $\eta = 10^{-3}$, and $\epsilon = 10^{-8}$. The learning rate was decayed in a step-wise manner by multiplying η with a factor of 0.1 after 3 consecutive epochs with no improvement in loss value on the validation dataset.

Similarly, we employed an early stopping scheme by monitoring the loss on the validation dataset and stopping the model training after 10 epochs of no improvement on $\mathcal{D}_{\text{EVAL}}$.

We tested four types of models in two categories: the first is a default split-stream model as shown in Figure 4.5 with and without weight decay (splitstream, splitstream-wd). The second is a variation of the split-stream model, but where the ψ_{clf} and ψ_{loc} modules are realized using depth-wise convolutions, such that each attention group is used only for that type of event. The second category is also tested with and without weight decay (splitstream-dw, splitstream-dw-wd).

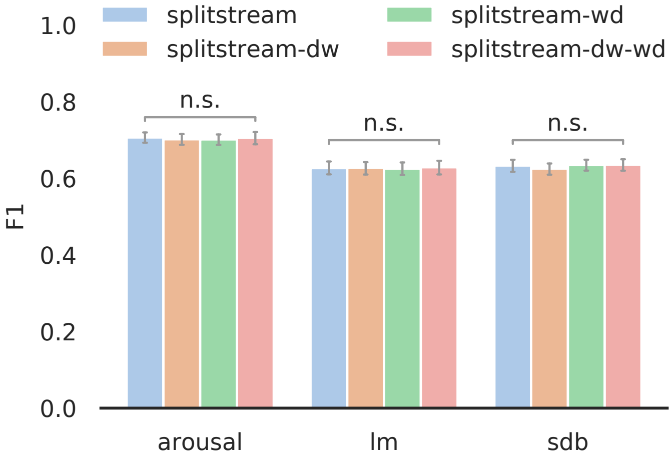


Figure 4.6: Architecture optimization.

4.3.2.13 *Performance evaluation*

Performance was quantified using precision, recall and F1 scores. Statistical significance in F1 score between groups was assessed with Kruskall-Wallis *H*-tests. The performance of joint vs. single-event detection models was tested with Wilcoxon signed rank tests for matched samples. The relationships between true and predicted ArI, AHI, and LMI were assessed using linear models and Pearsons r^2. Significance was set at $\alpha = 0.05$.

4.3.3 *Results and discussion*

4.3.3.1 *Model architecture evaluation*

We found no significant differences in F1 performance for either Ar (Kruskal-Wallis $H = 0.961$, $p = 0.811$), LM ($H = 0.230$, $p = 0.973$), or SDB detection ($H = 2.838$, $p = 0.417$), when evaluating the model architectures on $\mathcal{D}_{\text{EVAL}}$. Based on this result, all further modeling was based on the default splitstream architecture for simplicity.

4.3.3.2 *Joint vs. single event detection*

For each event type, we evaluated the F1 score as a function of classification threshold θ on $\mathcal{D}_{\text{EVAL}}$ for both the joint detection model as well as the single-event models. It can be observed in Figure 4.7 that for all three events, the joint detection model achieves higher F1 score, although the apparent increase is not as large for LM detection. This was also observed when evaluating the joint and single detection models with optimized thresholds on $\mathcal{D}_{\text{TEST}}$ for both Ar (Wilcoxon $W = 30440.0$, $p = 2.481 \times 10^{-127}$), LM ($W = 101103.0$, $p = 6.454 \times 10^{-60}$), and SDB detection ($W = 93647.0$, $p = 2.378 \times 10^{-64}$). Precision, recall and F1 scores for optimized models evaluated on $\mathcal{D}_{\text{TEST}}$ are shown in Table 4.5. These findings are interesting, because they provide evidence that the presence of different event types can module the detection of others, and that this can be modeled using automatic methods. This is in line with what previous studies have found e. g. on event-by-event scoring agreement in arousals, which improved significantly from 0.59 % to 0.91 %, when including respiratory signals in the analysis [71].

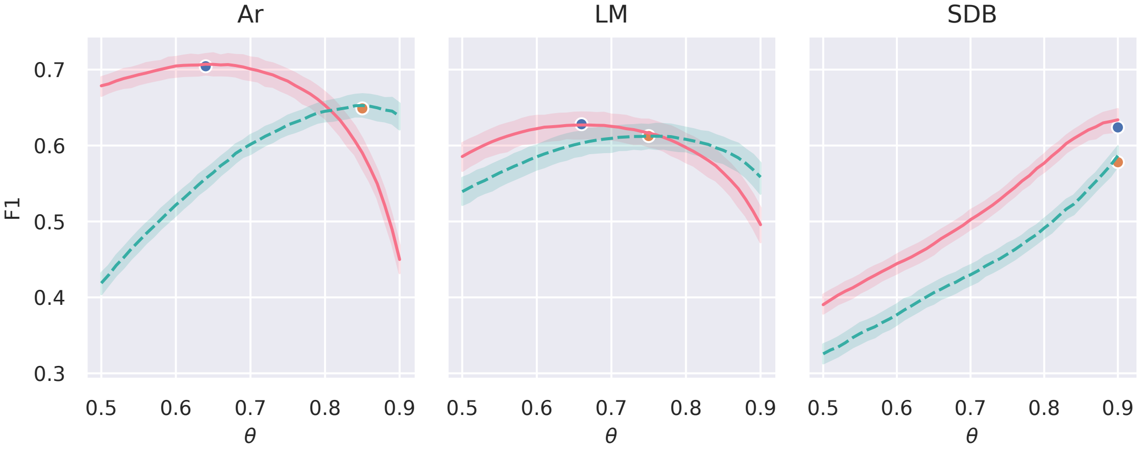


Figure 4.7: Optimizing F1 performance on $\mathcal{D}_{\text{EVAL}}$ as a function of θ). Full lines correspond to the joint model and dashed lines are the corresponding single-event detection model. The blue and orange dots correspond to optimized model performance on $\mathcal{D}_{\text{TEST}}$.

Table 4.5: Performance scores for optimized models evaluated on $\mathcal{D}_{\text{TEST}}$.

Event	Model	Precision	Recall	F1
Ar	Joint	0.759 ± 0.114	0.672 ± 0.125	0.704 ± 0.106
	Single	0.777 ± 0.107	0.571 ± 0.127	0.649 ± 0.113
LM	Joint	0.650 ± 0.169	0.647 ± 0.120	0.628 ± 0.123
	Single	0.661 ± 0.166	0.607 ± 0.116	0.613 ± 0.116
SDB	Joint	0.817 ± 0.142	0.526 ± 0.146	0.624 ± 0.115
	Single	0.765 ± 0.142	0.486 ± 0.121	0.578 ± 0.097

Metrics are shown aggregated across PSGs. Ar: arousal; LM: limb movement; SDB: sleep disordered breathing.

4.3.3.3 *Detection vs. manual scorings*

For each event type, we computed the correlation coefficient between the predicted and true index values (arousal index, ArI; apnea-hypopnea index, AHI; limb movement index, LMI), which is shown in Figure 4.9. We found a large positive correlation between true and predicted values for ArI ($r^2 = 0.73$, $p = 2.5 \times 10^{-285}$), AHI ($r^2 = 0.77$, $p = 9.3 \times 10^{-316}$), and LMI ($r^2 = 0.78$, $p = 3.1 \times 10^{-321}$).

A similar study using an automatic method for automatic detection of SDB and LM events found similar or higher correlations between automatic and manual scorings ($r^2 = 0.85$, and $r^2 = 0.79$, respectively), although their findings were based on almost 5 times as much data [89].

The authors pooled obstructive, central, mixed apneas, and 4% hypopneas into one category, apnea.

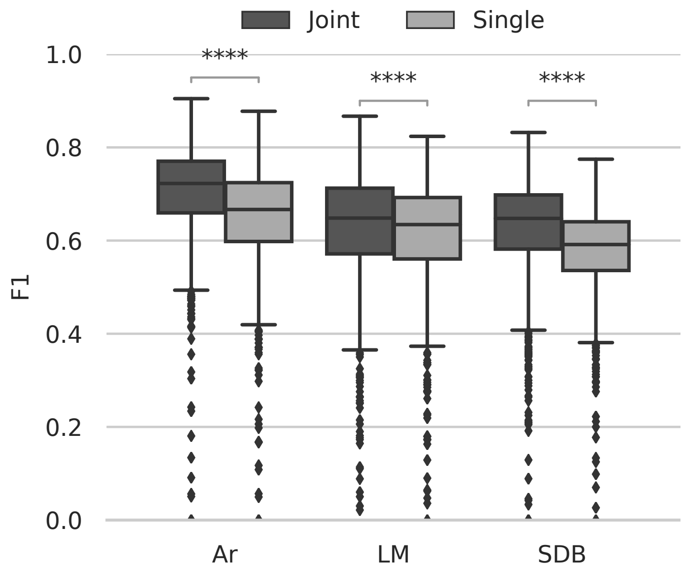


Figure 4.8: Evaluating optimized joint and single-event detection models on $\mathcal{D}_{\text{TEST}}$. ****: $p < 10 \times 10^{-4}$. Ar: arousal; LM: limb movement; SDB: sleep disordered breathing.

4.3.3.4 *Temporal characteristics*

We compared the temporal precision between manual and automatic event scoring by looking at the errors in onset (Δonset), offsets (Δoffset), and durations (Δdur.) calculated as

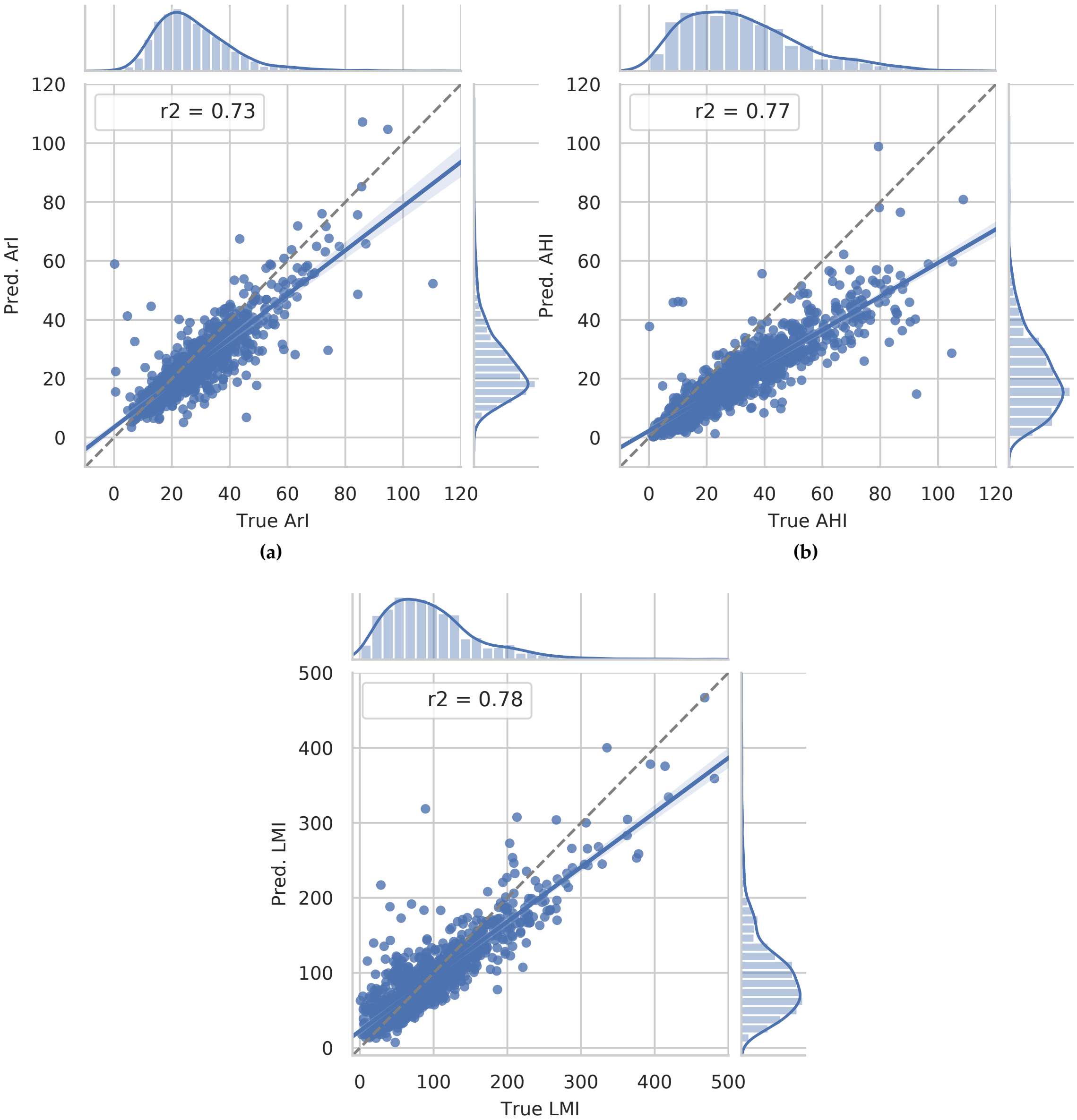
$$\Delta\,\text{onset} = \text{onset}_{\text{automatic}} - \text{onset}_{\text{manual}} \tag{4.21}$$

$$\Delta\,\text{offset} = \text{offset}_{\text{automatic}} - \text{offset}_{\text{manual}} \tag{4.22}$$

$$\Delta\,\text{dur} = \text{dur}_{\text{automatic}} - \text{dur}_{\text{manual}} \tag{4.23}$$

so that positive values of $\Delta\,\text{onset}, \Delta\,\text{offset}$ corresponds to a positive shift to the right (delayed prediction), and positive values of $\Delta\,\text{dur.}$ meaning an overestimation of the event duration compared to manual scoring. This is shown in Figure 4.10, where the blue distributions are the joint detection model for each event type, and the orange distributions are the corresponding single-event models. The distributions are shown as kernel density estimates superimposed on a histogram. For Ar events, the model overestimates the duration on average by a couple of seconds, which is caused by an earlier prediction of onset and delayed prediction of termination. For LM events, the model underestimates the duration by about half a second on average, which is due to earlier prediction of termination. For SDB events, the model overestimates the duration by about 25 seconds on average, which is caused by an earlier prediction of onset and delayed prediction of termination. These errors in predicted durations reflects the temporal characteristics of these events; LMs are shorter events, and it is thus unlikely to be overestimated by several seconds, while SDBs are longer events by one to two orders of magnitude, which also increases the size of the errors. Ars events are intermediate in length compared to LMs and SDBs, which is reflected in the error distributions.

Between 0.5 s *to* 10 s *per definition.*



(a)

(b)

(c)

Figure 4.9: Pearson correlation plots for each event type index between true and predicted values. The linear relationship is indicated with solid blue with 95% confidence intervals in light blue. Grey dashed lines indicate perfect correlation lines. ArI: arousal index; AHI: apnea-hypopnea index; LMI: limb movement index.

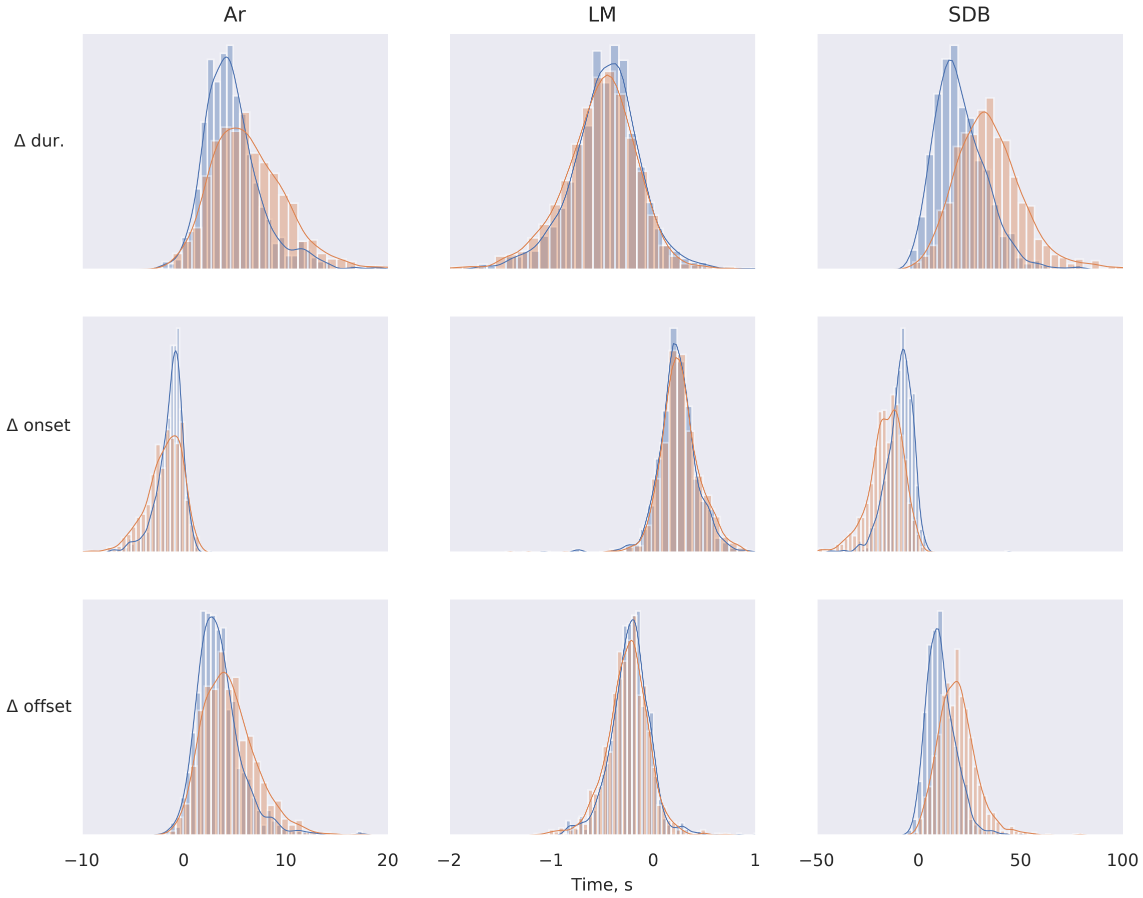


Figure 4.10: Temporal error metrics distributions across all events and PSGs. Positive values of Δonset, Δoffset means delayed predictions, while positive values of Δdur. means to an overestimation of event duration. Blue distributions are joint detection models, while orange distributions are the corresponding single-event models. Distributions are shown as kernel density estimates superimposed on a histogram. Ar: arousal; LM: limb movement; SDB: sleep disordered breathing.

4.4 PAPER V: DEEP TRANSFER LEARNING FOR IMPROVING SINGLE-EEG AROUSAL DETECTION

ABSTRACT: Datasets in sleep science present challenges for machine learning algorithms due to differences in recording setups across clinics. We investigate two deep transfer learning strategies for overcoming the channel mismatch problem for cases where two datasets do not contain exactly the same setup leading to degraded performance in single-EEG models. Specifically, we train a baseline model on multivariate polysomnography data and subsequently replace the first two layers to prepare the architecture for single-channel electroencephalography data. Using a fine-tuning strategy, our model yields similar performance to the baseline model (F1=0.682 and F1=0.694, respectively), and was significantly better than a comparable single-channel model. Our results are promising for researchers working with small databases who wish to use deep learning models pre-trained on larger databases.

4.4.1 *Methods*

NOTATION We denote by $[\![a, b]\!]$ the set of integers $\{n \in \mathbb{N} \mid a \leqslant n \leqslant b\}$ with $[\![N]\!]$ being shorthand for $[\![1, N]\!]$, and by $n \in [\![N]\!]$ the nth sample in $[\![N]\!]$. A model for a given experiment is denoted by $\mathcal{M}_{(\cdot)}$, while an optimized model is superscripted with a star as $\mathcal{M}^*_{(\cdot)}$. A segment of PSG data is denoted by $\mathbf{x} \in \mathbb{R}^{C \times T}$, where C, T is the number of channels and the duration of the segment in samples, respectively.

4.4.1.1 *Data*

We collected PSGs from 1500 subjects in the Osteoporotic Fractures in Men Sleep Study [103]–[105] from the NSRR [107], [108]. From each PSG, we extracted left and right EEG, left and right EOG, and chin EMG. EEG and EOG channels were referenced to the contralateral mastoid process. For each PSG, we also extracted time-stamped arousal scorings containing starts and durations of scored arousal events. We did not exclude any PSGs from this study based on sleep duration, number of arousal events, or similar criteria.

4.4.1.2 *Data partitioning*

The 1500 PSGs were initially partitioned into three subsets TRAIN$_1$, EVAL$_1$, and TEST$_1$ containing 400, 100 and 1000 PSGs, respectively. Furthermore, we additionally partitioned TEST$_1$ into three smaller subsets TRAIN$_2$, EVAL$_2$, and TEST$_2$ containing 400, 100, and 500 PSGs, respectively.

4.4.1.3 *Preprocessing pipeline*

All signals were resampled to 128 Hz using poly-phase filtering with a Kaiser window ($\beta = 5.0$) prior to subsequent processing. Extracted EEG and EOG signals were filtered with 2nd order Butterworth IIR bandpass filters with cutoff frequencies 0.3 Hz and 35 Hz. Chin EMG was filtered with a 4th order

Butterworth IIR highpass filter with a cutoff frequency of 10 Hz. Filtered signals were subsequently standardized by

$$\mathbf{x}^{(i)} = \frac{\tilde{\mathbf{x}}^{(i)} - \boldsymbol{\mu}^{(i)}}{\boldsymbol{\sigma}^{(i)}}, \tag{4.24}$$

where $\tilde{\mathbf{x}}^{(i)} \in \mathbb{R}^{C \times T}$ is the raw matrix containing C input channels and T samples, and $\boldsymbol{\mu}^{(i)}, \boldsymbol{\sigma}^{(i)} \in \mathbb{R}^{C}$ are the mean and standard deviation vectors for the *i*'th PSG, respectively.

The vectors contain the mean and standard deviation for each signal modality.

4.4.1.4 *Model setup*

We expand upon previous work using similar models for sleep event detection [113]–[115]. Briefly, the model takes as input a tensor of PSG data $\mathbf{x} \in \mathbb{R}^{C \times T}$ and outputs

$$\mathbf{z} = (\mathbf{p}, \mathbf{y}) \in \mathbb{R}^{N_d \times T' \times K} \times \mathbb{R}^{N_d \times T' \times 2} \tag{4.25}$$

containing predicted arousal probabilities $\mathbf{p}$ and associated starts and durations for predicted arousal events $\mathbf{y}$. The differentiable function underlying the model comprises a deep neural network architecture consisting of the following modules:

INPUT MIXING MODULE Here, non-linear combinations of the input PSG data $\mathbf{x}$ are made using a non-linear mixing block $\phi_{\text{mix}} : \mathbb{R}^{1 \times C \times T} \to \mathbb{R}^{C \times 1 \times T}$. This is implemented using single-strided 2D convolution operations with C $(C, 1)$-dimensional kernels.

FEATURE EXTRACTION MODULE This module contains two components. The first is a convolutional feature extraction block $\varphi_{\text{conv}} : \mathbb{R}^{C \times 1 \times T} \to \mathbb{R}^{f' \times 1 \times T'}$ consisting of $k_{\max}$ successions of convolutional, batch normalization, and rectified linear unit (ReLU) layers. The second is a recurrent feature extraction block $\varphi_{\text{rec}} : \mathbb{R}^{f' \times 1 \times T'} \to \mathbb{R}^{f' \times 2 \times T'}$ with $f' = f_0 2^{k_{\max}}$ hidden units. The φ_{conv} block is responsible for bulk feature extraction and temporal decimation using strided convolutions, while φ_{rec} processes the raw features across the reduced temporal dimension using a bidirectional gated recurrent unit with f' hidden units [117].

EVENT DETECTION MODULE The output from φ_{rec} is processed by two separate blocks: $\psi_{\text{clf}} : \mathbb{R}^{f' \times 2 \times T'} \to \mathbb{R}^{KN_d \times 1 \times T'}$ outputs the tensor $\mathbf{p}$ containing predicted arousal probabilities for each time point $t \in [\![T']\!]$ for each default event window. $\psi_{\text{loc}} : \mathbb{R}^{f' \times 2 \times T'} \to \mathbb{R}^{2N_d \times T'}$ outputs the tensor $\mathbf{y}$ containing predicted start time and durations of arousal events. Both ψ_{clf} and ψ_{loc} are implemented using $(2, 1)$ convolutions rather than convolutions over the entire volume as in [113]–[115]. This serves a dual purpose: reducing the number of parameters to make the network more memory-efficient and allowing the kernel and feature maps to be temporally invariant.

For a detailed description of the network architecture, see Table 4.6.

4.4.1.5 *Loss objective*

The network parameters were optimized according to a three-component loss objective comprising a localization loss ℓ_{loc} and a positive and negative classification loss ℓ_{+} and ℓ_{-}, respectively, such that

Loss function is another term.

$$\ell = \ell_{\text{loc}} + \ell_{+} + \ell_{-}. \tag{4.26}$$

Table 4.6: Network architecture overview.

	Layer type	Kernel	Stride	Filters	Input size	Output size	Activation
$\mathbf{x}$	Input	—	—	—	$C \times T$	$1 \times C \times T$	—
ϕ_{mix}	2D conv	$(C, 1)$	$(1, 1)$	C	$1 \times C \times T$	$C \times 1 \times T$	ReLU
$\varphi_{\text{conv}}^{(1)}$	2D conv	$(1, c)$	$(1, s)$	$2f_0$	$C \times 1 \times T$	$2f_0 \times 1 \times T/s$	—
	BN	—	—	$2f_0$	$2f_0 \times 1 \times T/s$	$2f_0 \times 1 \times T/s$	ReLU
$\varphi_{\text{conv}}^{(k)}$	2D conv	$(1, c)$	$(1, s)$	$f_0 2^k$	$f_0 2^{k-1} \times 1 \times T/s^{k-1}$	$f_0 2^k \times 1 \times T/s^k$	—
	BN	—	—	$f_0 2^k$	$f_0 2^k \times 1 \times T/s^k$	$f_0 2^k \times 1 \times T/s^k$	ReLU
φ_{rec}	bGRU	—	—	f'	$f' \times 1 \times T'$	$f' \times 2 \times T'$	—
ψ_{clf}	2D conv	$(2, 1)$	$(1, 1)$	KN_d	$f' \times 2 \times T'$	$KN_d \times 1 \times T'$	Softmax
ψ_{loc}	2D conv	$(2, 1)$	$(1, 1)$	$2N_d$	$f' \times 2 \times T'$	$2N_d \times 1 \times T'$	Linear
$\mathbf{z}$	Output, $\mathbf{p}$	—	—	—	$KN_d \times 1 \times T'$	$N_d \times T' \times K$	—
	Output, $\mathbf{y}$	—	—	—	$2N_d \times 1 \times T'$	$N_d \times T' \times 2$	—

$\mathbf{x}$, input containing PSG data; $\mathbf{z}$, output containing predicted arousal probabilities and associated start and duration predictions; ϕ_{mix}, non-linear mixing block; φ_{conv}, convolutional feature extraction block, $k \in [\![2, k_{\text{max}}]\!]$; φ_{rec} recurrent feature extraction block; ψ_{clf}, event classification block; ψ_{loc}, event localization block; C, number of input channels; T, number of samples in a segment of PSG data; c, temporal kernel size; s, temporal stride; f_0, base number of feature maps; $f' = f_0 2^{k_{\text{max}}}$, maximum number of feature maps; $T' = T/s^{k_{\text{max}}}$, reduced temporal dimension in samples; N_d, number of default event windows in segment; K, number of classes; ReLU: rectified linear unit; bGRU: bidirectional gated recurrent unit; BN: batch normalization.

The localization loss was calculated using a Huber function

$$\ell_{\text{loc}} = \frac{1}{N_{\pi \setminus \emptyset}} \sum_{i \in \pi \setminus \emptyset} h^{(i)} \tag{4.27}$$

$$\mathbf{h} = \begin{cases} 0.5(\mathbf{y} - \mathbf{t})^2, & \text{if } |\mathbf{y} - \mathbf{t}| < 1, \\ |\mathbf{y} - \mathbf{t}| - 0.5, & \text{otherwise,} \end{cases} \tag{4.28}$$

where $i \in \pi \setminus \emptyset$ indicates event windows with a non-empty arousal target. Contributions from the positive/negative classification losses were calculated using a focal loss function [177]:

$$\ell_+ = \frac{1}{N_{\pi \setminus \emptyset}} \sum_{i \in \pi \setminus \emptyset} -\alpha(1 - \mathbf{p})^\gamma \log(\mathbf{p}), \text{ and} \tag{4.29}$$

$$\ell_- = \frac{1}{N_{\pi = \emptyset}} \sum_{i \in \pi = \emptyset} -\alpha(1 - \mathbf{p})^\gamma \log(\mathbf{p}), \tag{4.30}$$

where $\alpha = 0.25$ and $\gamma = 2$. This serves to counter the class imbalance in a single data segment, which typically consists of many event windows with few positive examples.

4.4.1.6 *Experimental setups*

We investigated the channel mismatch problem with the following four experimental setups:

FULL MONTAGE BASELINE (FM) In this experiment, we trained the event detection algorithm on TRAIN$_1$ using $C = 5$ channels: left/right central EEG, left/right EOG, and chin EMG. Convergence and the optimal detection threshold were assessed on EVAL$_1$ and performance was evaluated on TEST$_2$. The optimal baseline model was used as an initialization for the two transfer learning experiments described below.

PRETRAINING (PT) The optimal model $\mathcal{M}^*_{\text{FM}}$ was used in this experiment as an initialization for $\mathcal{M}_{\text{PT}}$. We adjusted the mixing module and first convolutional layer in the feature extraction module to account for the channel mismatch by replacing the convolutional and batch normalization layers, and subsequently trained these from scratch. The rest of the weights and bias terms were frozen to the optimized values from $\mathcal{M}^*_{\text{FM}}$. The network was trained on TRAIN$_2$ with only $C = 1$ channels (left central EEG, C3). Convergence and optimal detection thresholds were assessed on EVAL$_2$, while final performance was evaluated on TEST$_2$.

FINE-TUNING (FT) Similar to PT, the optimal model $\mathcal{M}^*_{\text{FM}}$ was used in this experiment as an initialization for $\mathcal{M}_{\text{FT}}$. Also, the mixing module and first convolutional layer in the feature extraction module were likewise adjusted. However, all other layers in $\mathcal{M}_{\text{FT}}$ were permitted to be further optimized by fine-tuning weights and bias terms during training. The model was trained using the same 400 PSGs from TRAIN$_2$ with the same $C = 1$ channel configuration as in PT.

SINGLE EEG BENCHMARK (SE) We benchmarked our two transfer learning experiments to a comparable situation in which an event detection model was trained on the same PSGs in TRAIN$_2$ using only the left central EEG (C3).

4.4.1.7 *Network optimization*

In all experimental runs, we optimized the loss objective in Equation (4.26) using the Adam optimization algorithm with a learning rate of $\alpha = 10^{-3}$ and the default parameter values $(\beta_1, \beta_2) = (0.9, 0.999)$ as suggested in [97]. We applied the same data sampling strategy as proposed in [115], in which a segment of data is sampled such that it contains at least 50% of a randomly sampled event across all PSGs. We used a default event window size of 15 s with 50 % overlap as this was found previously to work well for arousal detection [115].

All experiments were implemented in PyTorch 1.2 [166].

4.4.1.8 *Performance evaluation*

Bipartite matching was used to match detected and true events during training and testing. At test time, detected events were subjected to non-maximum suppression based on an IoU of at least 0.5 between detected and true events. We evaluated the performance of our experimental setups using precision, recall and F1 scores.

4.4.1.9 *Statistical analysis*

We used Kruskal–Wallis one-way analysis of variance tests for differences in performance metrics between groups (SE, FT and PT) with a significance

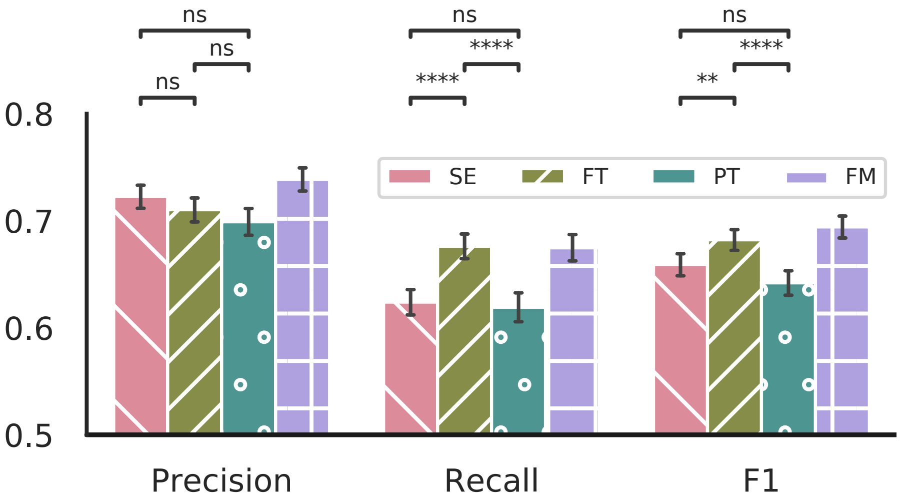


Figure 4.11: Performance metrics as evaluated on TEST$_2$ for each experimental setup. Metrics are shown as means with 95% confidence interval as error bars. Note the y-axis scaling. SE: single-EEG. FT: fine-tuning. PT: pre-training. FM: full montage. ns: not significant, **: $p_{adj} \leqslant 10^{-2}$; ****: $p_{adj} \leqslant 10^{-4}$.

Table 4.7: Performance metrics across experiments.

Experiment	Precision	Recall	F1
FM	0.739 ± 0.122	0.675 ± 0.139	0.694 ± 0.115
SE	0.723 ± 0.124	0.624 ± 0.137	0.659 ± 0.117
FT	$\mathbf{0.710 \pm 0.128}$	$\mathbf{0.676 \pm 0.130}$	$\mathbf{0.682 \pm 0.110}$
PT	0.699 ± 0.141	0.619 ± 0.153	0.642 ± 0.129

Metrics are shown evaluated on TEST$_2$ as means ± standard deviation. Best performing transfer learning experiment is shown in bold. SE: single-EEG. FT: fine-tuning. PT: pre-training. FM: full montage.

level of $\alpha = 0.05$. Post-hoc testing was performed with Mann-Whitney U-tests for each pair-combination (SE/FT, SE/PT, and FT/PT) likewise with $\alpha = 0.05$. We accounted for multiple comparisons by adjusting p-values with Bonferroni corrections.

4.4.2 *Results and discussion*

We present the results of the transfer learning experiments (FT, PT) as well as the baseline and benchmark experiments (FM, SE) in Figure 4.11 and Table 4.7. Performance metrics were not calculated for 10 subjects in TEST$_2$, as these did not have any scored arousals and are thus not reflected in Figure 4.11 and Table 4.7.

The baseline F1 performs slightly lower than previously reported (0.694 ± 0.115 vs. 0.749 ± 0.105 [115]). However, our baseline model was trained on 400 subjects compared to 1485 in [115], which would account for the lower F1 score. By reducing the available input channels from $C = 5$ different modalities to $C = 1$ EEG channel as in the SE benchmark experiment, the F1 score drops to 0.659 ± 0.117, while the precision and recall scores likewise drop from 0.739 ± 0.122 to 0.723 ± 0.124, and 0.675 ± 0.139 to 0.624 ± 0.137, respectively.

We found statistically significant differences in F1 scores between SE, FT, and PT ($p = 3.189 \times 10^{-7}$). Post-hoc testing further revealed statistically significant differences between SE and FT ($p_{adj} = 2.224 \times 10^{-3}$), and FT and PT ($p_{adj} = 2.685 \times 10^{-7}$), but not between SE and PT ($p_{adj} = 0.080$). We also found that recall scores differed between experimental setups ($p = 7.085 \times 10^{-13}$). Post-hoc testing showed statistically significant differences between SE, FT ($p_{adj} = 5.180 \times 10^{-11}$), and FT and PT ($p_{adj} = 1.440 \times 10^{-9}$), but not between SE and PT ($p_{adj} = 1.000$). Lastly, we saw statistically significant differences in precision scores between experimental setups ($p = 0.033$), subsequent post-hoc testing did not reveal any statistical significant differences, when adjusting for multiple comparisons using the Bonferroni procedure (SE/FT, $p_{adj} = 0.214$; FT/PT, $p_{adj} = 1.000$; SE/PT, $p_{adj} = 0.037$).

Our results show, that for some scenarios, we can learn and effectively transfer information present in multi-variate PSG data to a target domain containing only a single EEG channel. Specifically, the performance of our fine-tuning strategy is high enough that the mean F1 scores across subjects are statistically insignificant, when comparing FT and FM setups (not shown).

Previous related work focused on the channel mismatch problem, when comparing different, but the same number of, channel modalities such as transferring EEG-based models to EOG-based target domains, and thus did not investigate how changing the model architecture might impact performance [163], [164]. In this work, we investigated transfer learning when the source and target domains only overlap by one input channel. This necessitates changing some parts of the underlying model architecture to accommodate the different number of input channels, and these changes might impact downstream feature extraction. We did not explore simply zeroing out a large number of input channels in this work, as this requires exhaustive search of which channel indices to zero out in the model based on the number of target input channels.

Our study applied a simple optimization strategy for the transfer learning experiments, which might limit the potential performance gain. This is especially relevant for the FT experiment. For example, one could experiment with different learning rates and scheduling schemes for the initial layers and pre-trained layers, such that the initial layers were trained with a higher relative learning rate to compensate for their lack of initial training.

Furthermore, we explored transfer learning for the channel mismatch problem in a single cohort of patient recordings. Future directions of this research will investigate scenarios, where both the source and target domains, and the datasets are different.

4.5 CHAPTER SUMMARY

RH 2: *Advanced biomedical signal processing and machine learning algorithms can be used for efficient, high-performing analysis of sleep studies with regards to sleep events.*

RQ 2.1: *can sleep events be detected precisely and reliably using novel machine learning algorithms.*

RQ 2.3: *how can we overcome the channel mismatch problem for sleep event detection.*

RQ 2.2: *can the detection of one event class modulate the detection of an event from another class.*

This chapter concerned the detection of sleep events motivated by research hypothesis **RH 2**, and research questions **RQ 2.1**, **RQ 2.3**, and **RQ 2.2**.

The first study described the MSED model. Here, the main objective was to detect multiple sleep events in PSG studies simultaneously and independently using only a single model, which was trained and tested on 1485 and 1000 PSGs, respectively. Optimal arousal detection was obtained by including a recurrent neural network module and using a dynamic default event window yielding an F1 score of 0.75, while optimal leg movement detection was obtained with a static window yielding an F1 score of 0.65.

The second study in this chapter presented the application of the MSED algorithm for arousal detection under the channel mismatch problem. We investigated two deep transfer learning strategies for overcoming the channel mismatch problem for cases, where two datasets do not contain exactly the same setup leading to degraded performance in single-EEG models. Using a fine-tuning strategy, the model yielded similar performance to the baseline model (F1 = 0.68 and F1 = 0.69, respectively), and was significantly better than a comparable single-channel model. While these results are promising, they will have to be validated in a larger setting across separate cohorts.

Motivated by the preliminary results obtained previously, we investigated whether the MSED model could be extended with more input channels for added detection of sleep disordered breathing events, the results of which are shown in the last study of this chapter. We tested different variations on network architecture and found that a split-stream network with each stream responsible for separate sets of input channels was beneficial for our task. The results from evaluating on 1000 separate PSG recordings was F1 scores of 0.70, 0.63 and 0.62 for Ar, LM, and SDB detection, respectively. Interestingly, we also found that the model performed better with respect to F1 score for each separate class, when detecting events jointly instead of using single models for each class. However, this model remains severely understudied, and future efforts should concentrate on including more data sources, evaluating on a dataset containing multiple scorers, and benchmarking against state of the art methods.

5 CLASSIFICATION OF SLEEP DISORDERS

Break the cycle, Morty. Rise above. Focus on science.

— Rick Sanchez
Rick and Morty, season 1, episode 9

This chapter aims to build upon the knowledge and methods introduced previously by applying them in a clinical setting. Specifically, I will describe how we applied one of the sleep stage classification algorithms introduced in Chapter 3 to identify patients with narcolepsy, which is a sleep disorder characterized by a dysfunctional regulation of the sleep-wake switch described in Section 2.1.2.

The content of this chapter is based on the original publication

> J. B. Stephansen* *et al.*, "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy", *Nat. Commun.*, vol. 9, p. 5229, 2018. DOI: 10.1038/s41467-018-07229-3[1]

5.1 RESEARCH BACKGROUND

Sleep disorders and sleep dysregulation impact over 100 million Americans by contributing to a range of cardiovascular, metabolic and psychiatric disorders, such as obesity, diabetes, and depression. Generalized sleep deprivation also negatively impairs performance, judgment, and mood, and is a major preventable contributor to motor-vehicle-related accidents [15]. There are approximately 90 different sleep disorders currently recognized and described in the ICSD grouped into six categories: insomnias, circadian rhythm sleep-wake disorders, central hypersomnias (e. g. narcolepsy), sleep-related breathing disorders (e. g. obstructive sleep apnea), parasomnias (e. g. sleepwalking, RBD), and sleep-related movement disorders (e. g. periodic leg movement disorder (PLMD) and restless legs syndrome) [1].

Among these pathologies, NT1 is unique as a disorder with a known, discrete pathophysiology—a destruction of hypocretin neurons in the hypothalamus, which is most likely of autoimmune origin [178]–[180]. This is

[1] Creative Commons Attributes 4.0 International License: http://creativecommons.org/licenses/by/4.0/

reflected in the cerebrospinal fluid (CSF) concentrations of the hypocretin-1 neuropeptide, where a concentration below $110\,\mathrm{pg\,mL^{-1}}$ is considered indicative of narcolepsy [1].

Hypocretin-1 is also known as orexin-A. Although debated, there is also a narcolepsy type 2, which does not exhibit low CSF hypocretin levels [181].

Typically beginning in childhood or adolescence, narcolepsy affects approximately 0.03 % of the US, European, Korean and Chinese populations [64]. Unique to narcolepsy is the extremely strong association with the genetic marker HLA-DQB1*06:02 [182], and a well-characterized set of sleep disturbances that include short sleep latency, rapid transitions into REM sleep and poor nocturnal sleep consolidation. The pathology also includes episodes of *sleep-wake dissociations*, where the neuron groups in the sleep-wake or REM-NREM switches fire at the wrong time. This results in the clinical manifestations shown with parentheses in Figure 2.5.

The differentiation of sleep stages is also particularly important for the diagnosis of narcolepsy. Current diagnostic guidelines for NT1 require a full-night PSG and a multiple sleep latency test (MSLT) the following day, where patients are asked to nap 4 to 5 times for 20 min every 2 h during the daytime, and for each nap, the sleep latency and REM latency are noted [183]. A mean sleep latency (MSL) less than 8 min and the presence of at least 2 sleep onset REM periods (SOREMPs) during the MSLT, or 1 SOREMP plus a REM latency less than 15 min during nocturnal PSG are diagnostic criteria for NT1 [1]. In a recent large study of the MSLT, specificity and sensitivity for NT1 were 98.6 % and 92.9 % in comparing 516 NT1 versus 516 controls, respectively; and 71.2 % and 93.4 % in comparing 122 NT1 cases versus 132 other hypersomnia cases, respectively [111]. Similar sensitivities of 75 % to 90 % and specificities of 90 % to 98 % have been reported by others in large samples of hypersomnia cases versus NT1 [179], [184]–[187]. The MSLT is thus both highly specific and highly sensitive, making it incredibly valuable as a diagnostic tool.

A MSL less than 8 min is indicative of excessive sleepiness

A SOREMP is defined as REM latency less than 15 min following sleep onset in a nap.

5.1.1 *Research motivation and objectives*

In Section 3.4, we saw how a sleep stage classification algorithm could be constructed to reliably classify sleep stages as well or better than human experts. The results presented an interesting observation: sleep stage classification performance was unperturbed by existing sleep disorders, except in patients with narcolepsy. Furthermore, when comparing the hypnodensities in patients with and without narcolepsy, the former exhibited a much more diffuse sleep architecture with less pronounced sleep-wake cycles and increased REM/W/N1 disassociation. This is illustrated in Figure 5.1 where the bottom (top) trace shows a hypnodensity graph for a subject with (without) narcolepsy.

These findings motivated a novel research question with is directly associated with research hypothesis **RH 3**:

RH 3: *Advanced biomedical signal processing and machine learning algorithms can be used for efficient, high-performing analysis of sleep studies with regards to sleep disorders.*

RQ 3.1 based on a single overnight PSG recording, is it possible to diagnose narcolepsy with the same level of performance as the current clinical gold standard?

Derived from the research hypothesis and associated question, the following objectives were formulated:

(i) the model should be capable of diagnosing narcolepsy from the hypnodensity representation of a PSG study;

(ii) the model should have comparable or higher level of performance as the gold standard PSG-MSLT combination.

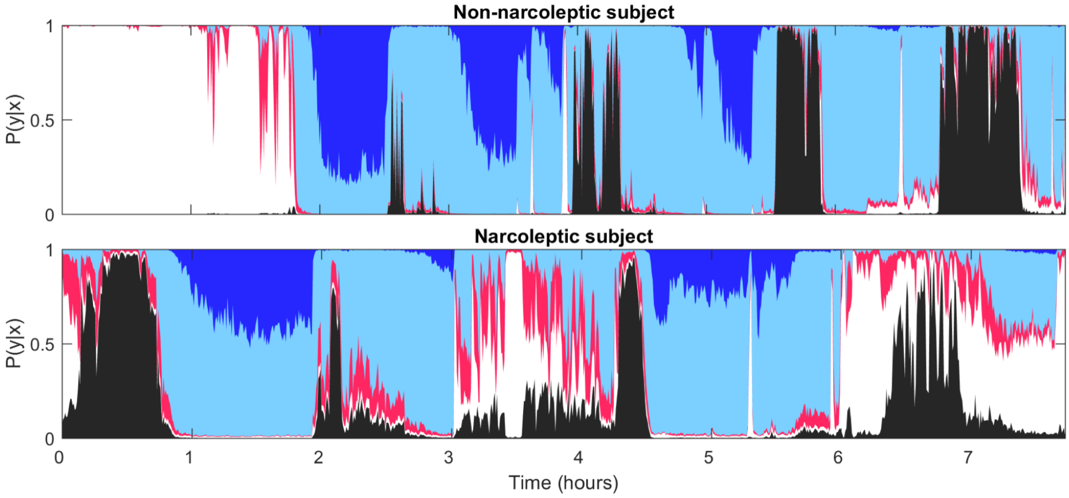


Figure 5.1: Examples of hypnodensity graph in subjects with and without narcolepsy. Hypnodensity for a subject without narcolepsy (top) and a subject with narcolepsy (bottom). Color codes: white, W; red, N1; light blue, N2; dark blue, N3; black, REM.

The following sections describe the steps taken to complete the posed objectives and answer the research question.

5.2 PAPER III: NEURAL NETWORK ANALYSIS OF SLEEP STAGES ENABLES EFFICIENT DIAGNOSIS OF NARCOLEPSY

ABSTRACT: Analysis of sleep for the diagnosis of sleep disorders such as NT1 currently requires visual inspection of polysomnography records by trained scoring technicians. Here, we used neural networks in approximately 3000 normal and abnormal sleep recordings to automate sleep stage scoring, producing a hypnodensity graph—a probability distribution conveying more information than classical hypnograms. A NT1 marker based on unusual sleep stage overlaps achieved a specificity of 96% and a sensitivity of 91%, validated in independent datasets. Addition of HLA-DQB1*06:02 typing increased specificity to 99%. Our method can reduce time spent in sleep clinics and automates NT1 diagnosis. It also opens the possibility of diagnosing NT1 using home sleep studies.

5.2.1 *Methods*

The following sections describe the narcolepsy model aspects in detail from initial hypnodensity computation, to feature engineering, and finally data modeling using Gaussian process (GP) classification algorithms. The main outcome of this approach is to be able to classify a hypnodensity representation of a PSG as either being positive or negative for NT1.

5.2.1.1 *Data descriptions*

PATIENT-BASED AUSTRIAN HYPERSOMNIA COHORT Patients in this cohort were examined at the Innsbruck Medical University in Austria as described in Frauscher *et al.* [188]. The AHC contains 118 PSGs in 86 high pretest probability patients for narcolepsy (see Table 3.11 for details). 42 patients (81 studies) are clear T1N with cataplexy cases, with all but 3 having a positive MSLT (these three subjects had a mean sleep latency (MSL)>8 minutes but multiple SOREMPs). The rest of the sample has idiopathic hypersomnia and type 2 narcolepsy. Four patients have an AHI>15/hour and 25 had a PLMI>15/hour. Almost all subjects had two sleep recordings performed, which were kept together such that no two recordings from the same subject were split between training and testing partitions.

THE JAZZ CLINICAL TRIAL SAMPLE This sample includes seven baseline sleep PSGs from five sites taken from a clinical trial study of sodium oxybate in narcolepsy (SXB15 with 45 sites in Canada, USA, and Switzerland) conducted by Orphan Medical, now named Jazz Pharmaceuticals. The few patients included are those with clear and frequent cataplexy (a requirement of the trial) that had no stimulant or antidepressant treatment at baseline [189]. All 7 subjects in this sample were used exclusively for training the narcolepsy biomarker algorithm.

PATIENT-BASED ITALIAN HYPERSOMNIA COHORT Patients in this high pretest probability cohort (see Table 3.11 for demographics) were examined at the IRCCS, Istituto delle Scienze Neurologiche ASL di Bologna in Italy as described in Pizza *et al.* [190]. The IHC contains 70 NT1 patients (58 %

Table 5.1: Description of the various cohorts included in this study and how they were used.

Cohort	**Age**, years	**BMI**, kg m^{-2}	**Sex**, %	**Train**	**Test**	**Replication**	**NT1**, %	**H**, %
WSC	59.7 ± 8.4	31.6 ± 7.1	53.1	170	116	None	0	0
SSC	45.4 ± 13.8	23.9 ± 6.5	59.4	139	112	None	11.6	1.8
KHC	29.1 ± 13.2	24.1 ± 4.3	58.6	87	71	None	45.8	54.2
AHC	34.5 ± 13.8	25.9 ± 4.9	54	42 (76)	44 (84)	None	52.3	47.7
JCTS	53.2 ± 9.8	31.0 ± 4.4	57.1	7	None	None	100	0
IHC	33.7 ± 17.6	-	56.7	87	61	None	47.3	50
DHC	33.4 ± 14.8	24.8 ± 4.9	50	79	None	None	26.6	48.1
FHC	28.8 ± 15.2	24.4 ± 8.1	59	None	None	122	51.6	18
CNC	28.5 ± 16.9	23.2 ± 11.5	51.3	None	None	199	34.2	0
Total subjects				611	404	321		
Total PSGs				645	444	321		

WSC: Wisconsin Sleep Cohort; SSC: Stanford Sleep Cohort; KHC: Korean Hypersomnia Cohort; AHC: Austrian Hypersomnia Cohort; JCTS: Jazz Clinical Trial Sample; IHC: Italian Hypersomnia Cohort; DHC: Danish Hypersomnia Cohort; FHC: French Hypersomnia Cohort; CNC: Chinese Narcolepsy Cohort; NT1: narcolepsy type 1; H, unspecific hypersomnia (narcolepsy type 2 (NT2) and idiopathic hypersomnia).

male, 29.5 ± 1.9 years old), with either documented low CSF hypocretin levels (59 cases, all but 2 HLA DQB1*06:02 positive), or clear cataplexy, positive MSLTs and HLA positivity (11 subjects). As non-NT1 cases with unexplained daytime somnolence, the cohort includes 77 other patients: 19 with idiopathic hypersomnia, 7 with type 2 narcolepsy and normal CSF hypocretin-1, 48 with a subjective complaint of excessive daytime sleepiness not confirmed by MSLT, and 3 with secondary hypersomnia. Subjects in this cohort were used for training (n=87) and testing (n=61) the narcolepsy biomarker algorithm.

PATIENT-BASED DANISH HYPERSOMNIA COHORT Patients in this cohort were examined at the Rigshospitalet, Glostrup, Denmark as described in Christensen *et al.* [191]. The DHC contains 79 PSGs in controls and patients (see Table 3.11 for details). Based on PSG, multiple sleep latency test and cerebrospinal fluid hypocretin-1 measures, the cohort includes healthy controls (19 subjects), patients with other sleep disorders and excessive daytime sleepiness (20 patients with CSF hypocretin-1 ⩾ 110 pg/ml), narcolepsy type 2 (22 patients with CSF hypocretin-1 ⩾ 110 pg/ml), and T1N (28 patients with CSF hypocretin 1 ⩽ 110 pg/ml). All 79 subjects in this cohort were used exclusively for training the narcolepsy biomarker algorithm.

PATIENT-BASED FRENCH HYPERSOMNIA COHORT This cohort consists of 122 individual PSGs recorded at the Sleep-Wake Disorders Center, Department of Neurology, Gui-de-Chauliac Hospital, CHU Montpellier, France (see Table 3.11 for demographics). The FHC contains 63 subjects with T1N (all but two tested with CSF hypocretin-1 ⩽ 110 pg/ml, five below 18 years old, 55 tested for HLA, all positive for HLA DQB1*06:02) and 22 narcolepsy type 2 (19 with CSF hypocretin-1 > 200 pg/ml, and three subjects with CSF hypocretin-1 between 110 and 200 pg/ml, three HLA positive). The remaining 36 subjects are controls (15 tested for HLA, two with DQB1*06:02)

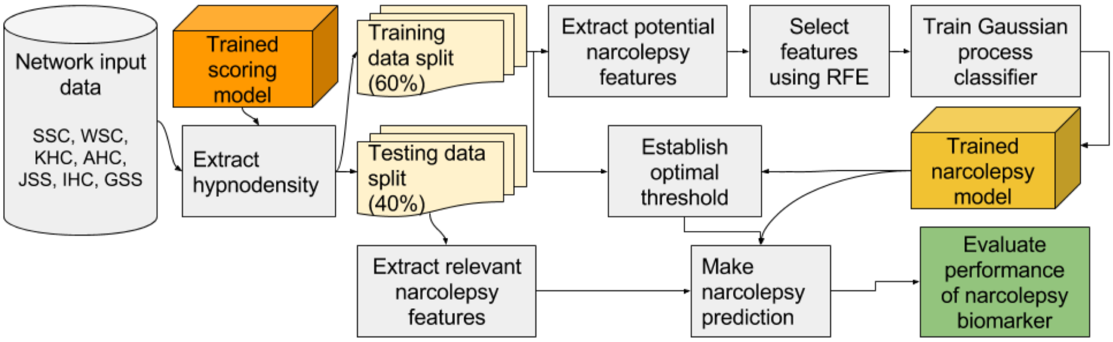

Network input data
SSC, WSC, KHC, AHC, JSS, IHC, GSS
Trained scoring model
Extract hypnodensity
Training data split (60%)
Testing data split (40%)
Extract potential narcolepsy features
Select features using RFE
Train Gaussian process classifier
Establish optimal threshold
Trained narcolepsy model
Extract relevant narcolepsy features
Make narcolepsy prediction
Evaluate performance of narcolepsy biomarker

Figure 5.2: Narcolepsy detector algorithm design. Hypnodensities are extracted from data, as described in Section 3.4. These data are separated into a training (60 %) and a testing (40 %) split. From the training split, 481 potentially relevant features described in Table 5.2 are extracted from each hypnodensity. The prominent features are selected using a recursive feature elimination (RFE) algorithm, and the narcolepsy detection model is trained using a GP model. The performance of the GP narcolepsy detection model is evaluated using the selected features computed on the test data.

without other symptoms of hypersomnia. The FHC was used as data for the replication study of the narcolepsy biomarker algorithm.

PATIENT-BASED CHINESE NARCOLEPSY COHORT This cohort contains 199 individual PSGs recorded (see Table 3.11 for demographics). The CNC contains 67 subjects diagnosed with T1N exhibiting clear-cut cataplexy (55 tested HLA DQB1*06:02 positive), while the remaining 132 subjects are randomly selected population controls (15 HLA DQB1*06:02 positive, 34 HLA negative, remaining unknown) [184]. Together with the FHC, the CNC was used as data for the replication study of the narcolepsy biomarker algorithm.

5.2.1.2 *Model overview*

The general pipeline for training and testing the narcolepsy model is shown in Figure 5.2. The input data sources from the cohorts described in Table 5.1 are shown on the left side. The PSGs from these cohorts are extracted and subjected to the sleep stage scoring model described in Section 3.4 resulting in a hypnodensity representation for each PSG.

The hypnodensity is further described in Section 3.4, but is essentially a probability distribution over sleep stages.

The hypnodensity data are split into training and testing subsets in a 60 %/40 % ratio. The training data are used for building the GP narcolepsy model with a subset of features determined using a feature reduction agorithm. Cross-validation was employed to determine the optimal classification threshold, and the performance of the classifier was determined on the held-out testing data.

5.2.1.3 *Feature extraction for NT1*

The following sections describe the features computed for each hypnodensity representation. Overall, the features fall into two categories: (i) features based on the dynamics between various stage combinations; and, (ii) features based on reported findings in the literature.

HYPNODENSITY-DERIVED FEATURES To quantify narcolepsy-like behavior for a single recording *i*, features were generated based on a proto-feature

Table 5.2: Description of each feature, how it is calculated, and how it is numerated.

#	Description	Formula
1	General prevalence of a value	$\log\left(\frac{1}{N}\sum_{n=1}^{N}\mathbf{\Phi}_n(\mathcal{S}_k)\right)$
2	Highest achieved value	$-\log(1-\max\mathbf{\Phi}_n(\mathcal{S}_k))$
3	Average fluctuations in value	$\log\left(\frac{1}{N}\sum_{n=1}^{N}\left\lvert\frac{d\mathbf{\Phi}_n(\mathcal{S}_k)}{dn}\right\rvert\right)$
4	Log of Shannon entropy	$\log\left(\frac{-\sum_i s_i^2\log s_i^2}{N}\right)$
5–8	Time until *p* times max. value	$\log\left(\text{first}_p\left(\frac{\text{cum sum}(\mathbf{\Phi}(\mathcal{S}_k))}{\text{sum}(\mathbf{\Phi}(\mathcal{S}_k))}\right)\times 30\right)$
9	Weighted maximum	$\sqrt{\max\mathbf{\Phi}(\mathcal{S}_k)\times\bar{\mathbf{\Phi}}(\mathcal{S}_k)}$
10	Weighted average fluctuation	$\left(\frac{1}{N}\sum_{n=1}^{N}\left\lvert\frac{d\mathbf{\Phi}_n(\mathcal{S}_k)}{dn}\right\rvert\right)\times\bar{\mathbf{\Phi}}(\mathcal{S}_k)$
11	Weighted Shannon entropy	$\log\left(\frac{-\sum_i s_i^2\log s_i^2}{N}\times\bar{\mathbf{\Phi}}(\mathcal{S}_k)\right)$
12–15	Weighted time until *p* max value	$\sqrt{\log\left(\text{first}_p\left(\frac{\text{cum sum}(\mathbf{\Phi})\mathcal{S}_k}{\text{sum}(\mathbf{\Phi}(\mathcal{S}_k))}\right)\times 30\right)}$

[4] The Shannon entropy is calculated using wavelet decompositions of $\mathbf{\Phi}(\mathcal{S}_k)$, where s_i contains the *i*th detail coefficient. This feature describes the amount of information contained in the signal.
[5–8] *p* here corresponds to 5 %, 10 %, 30 % and 50 %.
[12–15] *p* here corresponds to 5 %, 10 %, 30 % and 50 %.
Each individual feature is scaled by subtracting the mean and dividing by the difference between the 85th and 15th percentile values. Each value was assessed visually to ensure that the transformations and scaling was done optimally.

derived from *k*-combinations of $\mathcal{S} = \{\text{W}, \text{R}, \text{N1}, \text{N2}, \text{N3}\}$. For the *n*th 5, 15 or 30 s segment in recording *i*, a single *k*-combination is selected from the set of all *k*-combinations, and the proto-feature is then calculated as the sum of the pair-wise products of the elements in the single *k*-combination, such that

$$\mathbf{\Phi}_n^{(i)}(\mathcal{S}_k) = \sum_{\zeta\in[\mathcal{S}_k]^2}\prod_{s\in\zeta} p\left(s \mid \mathbf{x}_n^{(i)}\right), \quad p\in[0,1], \tag{5.1}$$

where $\mathbf{\Phi}_n^{(i)}$ is the proto-feature for the *n*th segment in recording *i*, $\zeta \in [\mathcal{S}_k]^2$ is a 2-tuple, or pair-wise combination, in the set of all pair-wise combinations in the *k*-combination of $\mathcal{S}$, and s is a single element, or sleep stage, in ζ. For $k \in [\![5]\!]$, there are 31 different $\mathcal{S}_k$, e.g. $\{\text{W}, \text{R}\}, \{\text{N1}, \text{N2}, \text{N3}\}$. The predicted probability of a 5, 15 or 30 s epoch belonging to a certain class in $\mathcal{S}$ given the data $\mathbf{x}_n^{(i)}$ is given by $p\left(s \mid \mathbf{x}_n^{(i)}\right)$. For every value of *k*, 15 features based on the mean, derivative, entropy and cumulative sum were extracted as shown in Table 5.2.

ADDITIONAL POLYSOMNOGRAM FEATURES Apart from the hypnodensity-derived features, we also defined features based on the conventional hypnogram analysis.

One set of such features was selected because they have been found to differentiate NT1 from other subjects in prior studies [192]–[196]. These include

- nocturnal REM sleep latency (REML) [111],
- presence of a nightly SOREMP with a REML less than 15 min [111],

- presence and number of SOREMPs during the night, where the SOREMPs are defined as REM sleep occurring after at least 2.5 min of either W or N1, and

- nocturnal sleep latency [192].

A short sleep latency is common in patients with NT1

Other features include

- a NREM fragmentation index defined as 22 or more occurrences, where sustained N2/N3 is broken by at least 1 min of N1/W [192], and

- the number of W/N1 hypnogram bouts longer than 3 min [192].

In this study we also explored:

- the cumulative W/N1 duration for wakefulness periods shorter than 15 min;

- cumulative REM duration following W/N1 periods longer than 2.5 min; and,

- total nightly SOREMP duration defined as the sum of REM epochs following 2.5 min W/N1 periods.

5.2.1.4 *Probabilistic models for diagnostic purposes*

Gaussian processes can also be viewed as a probabilistic extension of support vector machines.

The large set of features was reduced using a cross-validated RFE algorithm [197]. Using a threshold of 0.40 yielded 38 relevant features, which were fed to a GP classifier as described below. GP classifiers are non-parametric probabilistic models that produce robust non-linear decision boundaries using kernels and provide estimates of the uncertainties in classifications. This is useful when combining estimates, but also when making a diagnosis; if an estimate is particularly uncertain, a doctor may opt for more tests to increase certainty before making a diagnosis. During GP model building, a training dataset is used to optimize a set of hyper-parameters, which specify the kernel function, the basis function coefficients, here a constant, noise variance, and to form the underlying covariance and mean function from which inference about new cases are made [198]. In this case, the kernel is the squared exponential: Two classes were established: narcolepsy type 1 and "other", which contains every other subject. These were labeled 1 and -1 respectively, placing all estimates in this range. For more information on GP in general, see the textbook by Rasmussen and Williams[198], while more information on variational inference for scalable GP classification can be found in the paper by Hensman, Matthews, and Ghahramani[199] and by Matthews *et al.*[200].

HLA TESTING As described previously, 97 % of NT1 patients are HLA-DQB1*06:02 positive when the disease is defined biochemically by low CSF hypocretin-1, or by the presence of cataplexy coupled with clear MSLT findings [111], [182]. We implemented this feature as a binary-valued predictor resulting in negative narcolepsy predictions for subjects with a negative HLA-DQB1*06:02 test result.

HIGH PRETEST PROBABILITY SAMPLE MSLTs are typically performed in patients with daytime sleepiness that cannot be explained by OSA, insufficient sleep or circadian disturbances alone. These patients thus have a higher pre-test probability of having NT1 than random clinical patients and are then diagnosed with NT1 or NT2, idiopathic hypersomnia or subjective sleepiness based on MSLT results, cataplexy symptoms and human leukocyte antigen (HLA) results, if they are available. To test whether our detector differentiates NT1 from these other cases with unexplained sleepiness, we conducted a post-hoc analysis of the detector performance in these subjects extracted from both the test and replication datasets.

5.2.2 *Results*

The neural networks produce outputs that depend on evidence in the input data for or against a certain sleep stage based on features learned through training. We hypothesized that narcolepsy, a condition characterized by sleep-wake dissociation [190], [192], [201]–[203], would result in a greater than normal overlap between stages, such as that shown in Figure 5.1. Based on this result, we hypothesized that such sleep stage model outputs could be used as a biomarker for the diagnosis of NT1 using a standard nocturnal PSG rather than the PSG-MSLT combination.

To quantify narcolepsy-like behavior for a single recording, we generated features quantifying sleep stage dissociation using 16 sleep stage prediction models. These features were based on descriptive statistics and other features describing persistence of a set of new time series generated from every permutation product of the set of predicted sleep stages.

We also added features expected to predict narcolepsy based on prior work, such as REM sleep latency and sleep stage sequencing parameters. A RFE procedure was performed on extracted features with average outcome setting the optimal number of relevant features at 38 [197].

An optimal selection frequency cut-off of 0.40 was determined using a cross-validation setup on the training data. The selected features are described in Table 5.3 with detailed description of the eight most important features reported in Table 5.4.

That means including a feature if it was selected in 40 % of the cross-validation runs

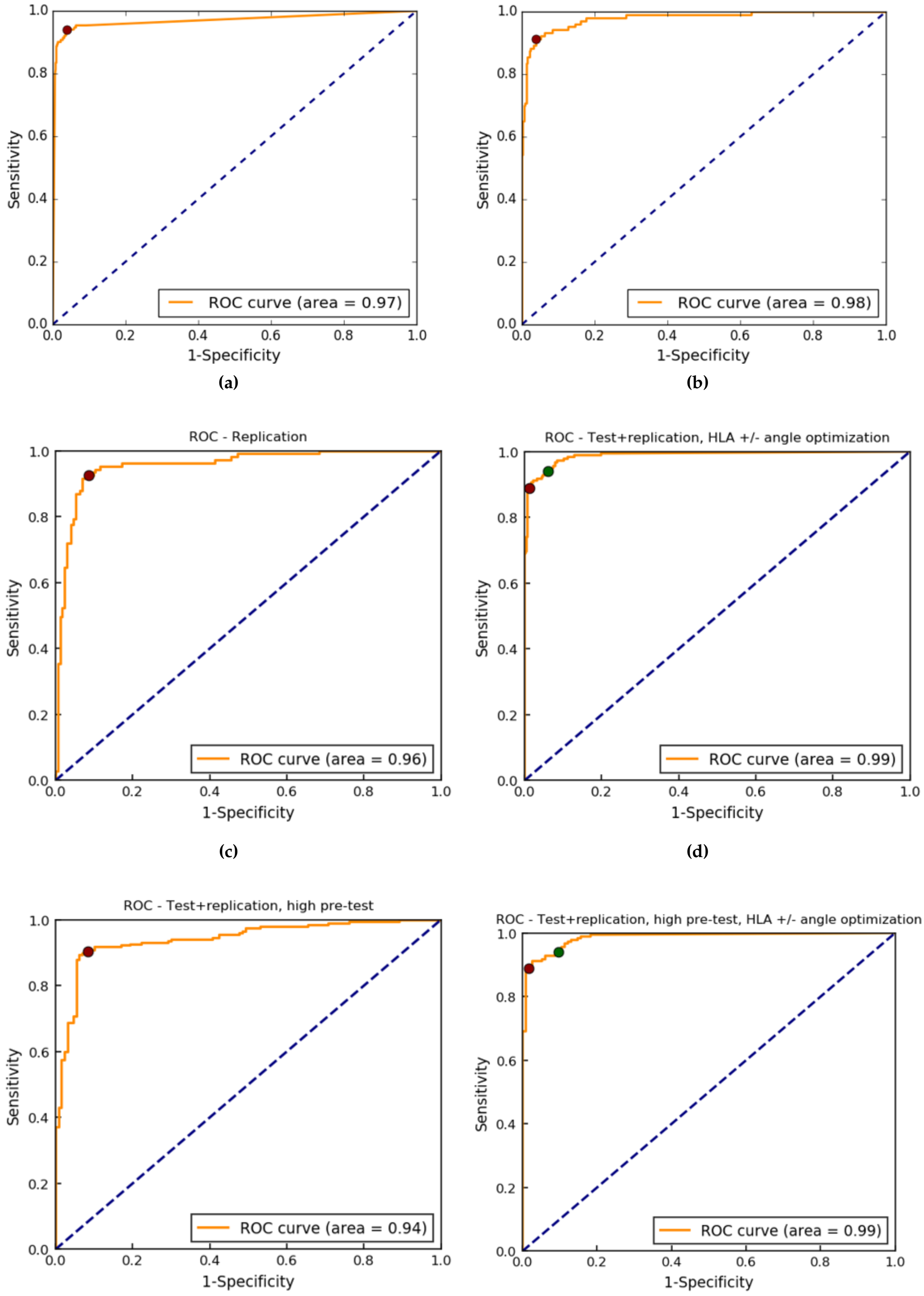
Final predictions were achieved by creating a separate GP narcolepsy classifier for each of the sleep scoring models used in the final implementation. The models were trained (tested) on data from seven (five) different cohorts with subsequent independent replication in two cohorts previously unseen by the algorithm, see Table 5.1. The algorithm produced values between -1 and 1, with 1 indicating a high probability of NT1. A cut-off threshold of -0.03 was determined using cross-validation on the training dataset, which is shown with red dots in Figure 5.3. This optimal trade-off achieves both high sensitivity and specificity, which is seen to translate well onto the test data and the replication sample in Figure 5.3b-c.

In the training data, a sensitivity of 94 % and specificity of 96 % was achieved, and in the testing data a sensitivity of 91 % and specificity of 96 % was achieved, while the sensitivity and specificity for the replication sample was 93 % and 91 %, respectively. When HLA was added to this model (Figure 5.3d–f), the sensitivity changed to 90 % and the specificity rose to 99 %, and the cut-off threshold was updated to −0.53 shown with green dots in Figure 5.3d–f. Furthermore, in the high pretest sample we obtained a sensitivity and specificity of 90 % and 92 %, which rose to 90 % and 98 % when adding HLA. More descriptive statistics including 95 % confidence intervals are shown in Table 5.5.

Table 5.3: Selection frequency and descriptions of each of the 38 features included in the Gaussian process model used for narcolepsy prediction. Numbers in second column correspond to feature number in Table 5.2.

	Feature	Stage combination	Frequency
1	12	W, N2, REM	1.00
2	Nightly SOREMPs		0.91
3	15	W	0.82
4	6	REM	0.82
5	2	W	0.68
6	2	N2, REM	0.68
7	14	W, N2	0.68
8	13	W, N1	0.64
9	5	N3	0.59
10	5	REM	0.59
11	13	N1, N2	0.59
12	8	N1	0.55
13	11	N1	0.55
14	7	W, N1, REM	0.55
15	5	W, N1, N3	0.55
16	6	W, N1, N3	0.55
17	1	W, N1, N2, REM	0.55
18	Hypnodensity sleep stage bout transitions from N2 to N3		0.55
19	Accumulation of W periods less than 15 min		0.50
20	Hypnodensity sleep stage bout transitions from W/N1 to REM		0.50
21	11	N3, REM	0.45
22	2	N1, REM	0.45
23	7	W, N2, N3	0.45
24	12	W	0.41
25	2	N1	0.41
26	12	N2	0.41
27	14	N2	0.41
28	7	N2, REM	0.41
29	8	N2, REM	0.41
30	6	N1, N2	0.41
31	15	N1, N2	0.41
32	15	W, N3	0.41
33	12	W, N1	0.41
34	5	W, N2, REM	0.41
35	1	W, N1, N3, REM	0.41
36	1	W, N1, N2, N3, REM	0.41
37	Accumulation of REM epochs following W periods		0.41
38	Hypnodensity sleep stage bout transitions from N2 to REM		0.41

SOREMP: sleep onset REM period; W: wakefulness; N1: non-rapid eye movement stage 1; N2: non-rapid eye movement stage 2; N3: non-rapid eye movement stage 3; REM: rapid eye movement.



(a) **(b)**

(c) **(d)**

(e) **(f)**

Figure 5.3: Diagnostic receiver operating characteristics curves for narcolepsy model displaying the trade-offs between sensitivity and specificity for the narcolepsy biomarker for (a) training sample, (b) testing sample, (c) replication sample, and (e) high pretest sample. (d)–(f) Adding HLA to model greatly increases specificity. Cut-off thresholds are presented for models with (red dot) and without HLA (green dot)

Table 5.4: Eight most frequently selected features for NT1 detection.

	Frequency	Description
1	1.00	Time until 5 % of the weighted sum of the product between W, N2, and REM calculated at every epoch has accumulated. This feature expresses the known sleep stage dissociation and altered sleep timing.
2	0.91	Number of SOREMPs appearing throughout the recording.
3	0.82	Time until 50 % of W in recording has accumulated weighted by total amount of W.
4	0.82	Shannon entropy of REM sleep. This expresses the amount of information held in a signal, or in this case, how many different values the REM sleep stage distribution obtains, i. e. how consolidated phases of REM are when the stage appears.
5	0.68	Maximum probability of W obtained in a recording.
6	0.68	Maximum value obtained of the product between N2 and REM probability in a recording.
7	0.68	Time until 30 % of the epoch-by-epoch sum product between W and N2 has accumulated, weighted by the sum total.
8	0.64	The time taken before 10 % of the epoch-by-epoch sum product between W and N1 has accumulated, weighted by the sum total.

NT1: narcolepsy type 1; W: wakefulness; N1: non-rapid eye movement stage 1; N2: non-rapid eye movement stage 2; N3: non-rapid eye movement stage 3; REM: rapid eye movement; SOREMP: sleep onset REM period.

Table 5.5: Descriptive statistics on the evaluation of the narcolepsy biomarker in models with and without the HLA biomarker. Mean value (top) and 95% confidence interval (bottom).

Model	**Accuracy**, %	**Sensitivity**, %	**Specificity**, %	**PPV**, %	**NPV**, %	**PSGs**	**NT1**, %
T	0.95	0.91	0.96	0.88	0.97	444	0.24
	0.92-0.97	0.84-0.96	0.93-0.98	0.80-0.93	0.95-0.99		
R	0.92	0.93	0.91	0.87	0.95	321	0.28
	0.88-0.95	0.87-0.97	0.87-0.95	0.80-0.93	0.92-0.98		
T+R, HLA	0.96	0.9	0.99	0.97	0.95	584	0.31
	0.94-0.97	0.84-0.93	0.98-1.00	0.94-0.99	0.93-0.97		
T+R, HLA,	0.94	0.94	0.94	0.88	0.97	584	0.31
optim.	0.92-0.96	0.90-0.97	0.92-0.96	0.83-0.92	0.95-0.99		
HPT, no HLA.	0.91	0.9	0.92	0.94	0.86	335	0.61
	0.87-0.94	0.86-0.94	0.86-0.96	0.91-0.97	0.80-0.91		
HPT, HLA	0.93	0.9	0.98	0.99	0.85	296	0.61
	0.90-0.95	0.84-0.93	0.96-1.00	0.97-1.00	0.79-0.91		
HPT, HLA,	0.93	0.94	0.9	0.94	0.9	296	0.61
optim.	0.90-0.95	0.90-0.97	0.85-0.95	0.90-0.97	0.85-0.95		

Performance on models with HLA typing is reported for both regular and optimized threshold, since the ROC curve changes by adding HLA. HLA: human leukocyte antigen; ROC: receiver operating characteristic; NT1: narcolepsy type 1; T, test dataset; R, replication dataset; HPT, high pre-test probability dataset; PPV, positive predictive value (precision); NPV, negative predictive value.

5.2.3 *Discussion*

Using our models, and considering how typical NT1 behaved in our sleep stage machine learning routines, we extracted features that could be useful to diagnose this condition.

Tables 5.3 and 5.4 reveal features found in nocturnal PSGss that discriminate NT1 from non-narcolepsy. One of the most prominent features, short latency REM sleep, bears great resemblance to the REM sleep latency, which is currently used clinically to diagnose narcolepsy. As short REML is calculated using fuzzy logic, it represents a latency where accumulated sleep suggests high probability of REM sleep occurrence. A short REM latency during PSG recording is extremely specific (99 %) and moderately sensitive (40 % to 50 %) for NT1 classification [111], [204]. The remaining selected features also describe a generally altered sleep architecture, particularly between REM sleep, light sleep and W. These dissociations mirror aspects of narcolepsy which are already known and thus reinforce their validity as biomarkers.

As opposed to a discrete REM latency scored by a technician

Short in this case typically means less than 15 min

Here light sleep is comprised of N1 and N2

For example, the primary feature, as determined by the RFE algorithm, was the time it took to reach5 % of the accumulated sum of the probability products between stages W, N2 and REM, which reflects the uncertainty between W, REM and N2 sleep at the beginning of the night. Specifically, for the *n*th epoch, the model will output probabilities for each sleep stage, and the proto-feature $\mathbf{\Phi}_n$ is calculated as

$$\mathbf{\Phi}_n = p(\textsc{w}) \times p(\textsc{n2}) + p(\textsc{w}) \times p(\textsc{r}) + p(\textsc{n2}) \times p(\textsc{r}) \tag{5.2}$$

The feature value is then calculated as the time it takes in minutes for the accumulated sum of $\mathbf{\Phi}_n$ to reach 5 % of the total sum $\sum_n \mathbf{\Phi}_n$. Since each probability product in $\mathbf{\Phi}_n$ reflects the staging uncertainty between each sleep stage pair, $\mathbf{\Phi}_n$ alone reflects the general sleep stage uncertainty for that specific epoch as predicted by the model. A high feature value is attained for epoch *n* when N2, W and REM are of equal probability and the remaining two sleep stages are close to zero. A PSG with a high staging uncertainty between sleep and wake early in the night would reach the 5 % threshold rapidly.

Using these features, we determined an optimal cut-off that discriminated narcolepsy from controls and other patients with specificity and sensitivity as high as the MSLT, notably when HLA typing is added. This is true for both the test and replication samples. Although we observed a small drop in specificity in the replication sample, the performance was similar to the MSLT when the efficacy of the detector was tested in the context of naive patients with hypersomnia in the high pretest probability sample.

See also Table 5.5 for details

Furthermore, MSLTs requires that patients spend an entire night and day in a sleep laboratory. This novel biomarker could allow for improved recognition of acNT1 cases at a reduced cost by only requiring a standard PSG screening as used for other sleep pathologies, such as OSA. A positive predictive value could also be provided depending on the nature of the sample and known narcolepsy prevalence. It also opens the possibility of using home sleep recordings for diagnosing narcolepsy. In this direction, because of the probabilistic and automatic nature of our biomarker, estimates from more than one night could be automatically analyzed and combined over time ensuring improved prediction. However, it is important to note that this algorithm will not replace the MSLT in the ability to predict excessive

This prevalence is low in general population screening, intermediary in a overall clinic population sample, and high in hypersomnia cohorts daytime sleepiness through the measure of mean sleep latency across daytime naps, which is an important characteristic of other hypersomnias.

When the staging data were presented as hypnodensity distributions, the model conveyed more information about the subject than through a hypnogram alone. This led to the creation of a biomarker for narcolepsy that achieved similar performance to the current clinical gold standard, the MSLT, but only requires a single sleep study. If increased specificity is needed, for example, in large-scale screening, HLA or additional genetic typing brings specificity above 99% without loss of sensitivity. This presents an option for robust, consistent, inexpensive and simpler diagnosis of subjects who may have narcolepsy, as such tests may also be carried out in a home environment.

This study shows how hypnodensity graphs can be created automatically from raw sleep study data, and how the resulting interpretable features can be used to generate a diagnosis probability for NT1. Another approach would be to classify narcolepsy directly from the neural network by optimizing the performance not only for sleep staging, but also for direct diagnosis by adding an additional softmax output, thereby creating a multitask classifier. This approach could lead to better predictions, since features are not then limited to by a designer imagination. A drawback of this approach is that features would no longer be as interpretable and meaningful to clinicians. If meaning could be extracted from these neural network generated features, this might open the door to a single universal sleep analysis model, covering multiple diseases. Development of such a model would require adding more subjects with narcolepsy and other conditions to the pool of training data.

5.3 chapter summary

This chapter concerned the use of signal processing and machine learning for the detection of sleep disorders, the topic of which directly relates to **RH 3**. Specifically, we were interested in the following question: based on a single overnight PSG recording, is it possible to diagnose narcolepsy with the same level of performance as the current clinical gold standard?

RH 3: *Advanced biomedical signal processing and machine learning algorithms can be used for efficient, high-performing analysis of sleep studies with regards to sleep disorders*

We designed a narcolepsy model using feature engineering and probabilistic machine learning models to classify NT1 patients with a 91 % sensitivity and 96 % specificity using a hypnodensity representation of a single, overnight PSG recording, which is the same level as the current gold standard. By adding HLA-DQB1*06:02 typing, the specificity increased to 99%.

Part III

OUTLOOK

6 DISCUSSION

Okay, well, sometimes science is more art than science, Morty. A lot of people don't get that.

— Rick Sanchez
Rick and Morty, season 1, episode 6

The objective of this thesis was to develop a system based on artificial intelligence, that can assist clinicians in the analysis of sleep studies. This was based on the hypothesis that advanced biomedical signal processing and machine learning algorithms can be used for efficient, high-performing analysis of sleep studies with regards to sleep stages, sleep events, and sleep disorders. In this thesis, the system was realized using three models. This chapter will touch on the results of three models as described in Chapters 3 to 5, and discuss aspects of including artificial intelligence in sleep clinics.

This thesis described methods for automatic sleep stage classification based on the manually annotated recordings according to the AASM guidelines. The outputs of this research theme were 1) the MASSC model, which yielded an accuracy of 84% and a Cohen's κ of 0.75 on a test set of 230 PSGs from WSC, which increased to an accuracy of 87% and Cohen's κ of 0.80 using data from five different cohorts; and, 2) the STAGES model, which yielded an accuracy of 87% and was found to be better than six independent scorers. The main question is whether we should train sleep stage classifiers to perform as well as one or several human eyes, or, if we should let the classifiers rely more on the hidden fluctuations in the PSG signals thereby allowing for more stages that are harder for the human eye to differentiate. This has been the subject of other research groups in the past years. For example, Stevner *et al.* modeled a combination of functional magnetic resonance imaging and EEG whole-brain dynamics using a hidden Markov model approach in 57 subjects [205]. Their main findings identified multiple distinct whole-brain network states, and highlight that individual sleep stages can be characterized by several of these latent states; e. g. they found that the W and N1 stages are increasingly heterogeneous comprising multiple latent states, while N2 and N3 comprise only a few. Similar findings have also been found in patients expressing insomnia and Parkinson's disease, where a topic model was constructed in both cases using latent Dirichlet allocation of "words" created from the EEG, EOG, and EMG, revealing that NREM could be comprised of several topics [99], [206], and that some individual topics could describe a dissociated sleep stage e. g. between N1 and N2 [100]. Further evidence also points towards that local areas in the brain can be in different sleep stages [207]. However, the clinical implications and utility of these findings are still unclear.

The same paper also found that individual signals possess different latent topics that describe various dissociations between sleep stages.

The second research theme concerned methods for automatic detection of sleep events focusing specifically on Ars, LMs, and SDB events. The output of this research theme was the MSED model, which yielded F1 scores of 0.704, 0.628, and 0.625 for Ar, LM, and SDB detection, respectively, when tested on 1000 PSGs. The MSED model was also used in a study concerning *transfer learning* in cases under the *channel mismatch problem*, where it was demonstrated that the F1 score could be recovered effectively using a fine-tuning strategy. The MSED model was not tested against multiple scorers, as some studies have already proposed for LM and Ar detection [158], [159], which makes direct comparison of final performance difficult. Indeed, this is true for any method based on artificial intelligence (AI), which has prompted the development of benchmark datasets and competitions for computer vision applications, such as object recognition and localization in the ImageNet challenge [122]. This would be an important step forward in the future of computational sleep science.

The third research theme on methods for sleep disorder detection was mainly focused on central hypersomnias, in particular the detection of NT1. The development of the STAGES algorithm, a machine learning model capable of classifying NT1 based on single-night PSG with high sensitivity and specificity even without the addition of HLA-DQB1*06:02 typing or hcrt serum levels, served as both the primary outcome and research contribution. A secondary contribution relates to the identification of several novel PSG features describing the dissociation between sleep stages, that was identified by the RFE procedure. Several research groups have found biomarkers describing increased sleep stage dissociation in narcolepsy [192], [201], [202], [208], Parkinson's disease [99], [100], [209], and insomnia [206], and the RFE procedure could very likely help in revealing new biomarkers for various diseases.

A well-known effect in sleep medicine is the first night effect, wherein subjects experience more W and less REM [19] during the first night of recording PSG, which is not seen in the second night of recording. Obviously, this will have an impact on the features extracted using the STAGES model, but the extent of this impact needs further research.

Apart from the PSG and HLA-DQB1*06:02 features, it is also possible that the addition of other types of research data would be of value. It could be theorized that the addition of questionnaire data such as the Narcolepsy Severity Scale [210] or the Alliance Sleep Questionnaire would be beneficial for example in distinguishing between hypersomnias. This has been explored in other studies where e. g. a digital sleep questionnaire was designed and used to classify and detect common societal sleep disturbances including insomnia, delayed sleep phase syndrome, insufficient sleep syndrome, and risk for obstructive sleep apnea [211].

Our model consistently was able to distinguish NT1 from NT2 as well as IH, but not NT2 from IH, indicating a clear distinction between the two narcolepsy types and that IH and NT2 might be more related than previously thought, which is also being considered by other research groups [181]. A panel consisting of European experts also recently reviewed the clinical findings for a future revision of the ICSD, in which they recommend three new categorizations of central hypersomnias as narcolepsy, idiopathic hypersomnia, and idiopathic excessive sleepiness [212].

The main findings of this thesis argue that clinical sleep medicine can benefit from incorporating computational methods such as deep learning and machine learning. However, major challenges still face the sleep science community from wholly adopting automatic sleep analysis methods: i) accessibility of curated data remains a major obstacle, which is especially important for those researchers interested in applying or developing machine learning algorithms and statistical methods, ii) how to share clinical data in a safe and regulated manner compliant with GDPR specifications. Several attempts have been made to address these issues; this thesis has relied heavily on data from the NSRR, which contain several high-quality databases from research cohorts and clinical trials in a stream-lined format [107], [108]. *PhysioNet* is another online resource containing vast amounts of freely accessible electrophysiological data, although the amount of sleep-related data is limited [79].

The AASM Artificial Intelligence in Sleep Medicine Committee recently published a statement on behalf of the AASM regarding the adoption and use of AI in the sleep clinics. Their official position of the AASM is that electrophysiological data such as those originating from PSG recordings, are well-suited for AI-based analysis due to the volume and variability of data. They argue, that AI analysis can add significant value by improving the efficiency of sleep labs, shifting focus away from manual analysis leading towards increased patient care, more rapid diagnosis and subsequent treatment [213]. However, they also emphasize that the goal of integrating AI-based systems into clinical practice should be to augment rather than replace expert-based evaluations, and that full-scale adoption is complicated by logistics, limited transparency of AI-based models, and ethical and regulatory issues [214].

The question then becomes if the field of sleep science and medicine is ready to move towards more automated sleep analysis? In response to this question, Lim *et al.* recently posed three advantages to automating sleep study scoring [215]. The first advantage point is that sleep clinics will have consistent results on the same sleep studies for both clinical and research purposes, which potentially can be used for building new treatment protocols and build upon the collective findings. The second mentioned advantage point is that automated sleep scoring will significantly reduce man-hours spent on PSG analysis, which will free up precious time for physicians and technicians to focus on patient care. The third and last mentioned advantage point is that automated sleep stage scoring has the potential to drive the field away from the current gold standard of scoring sleep stages and discover novel sleep stage characteristics in the brain.

Moving beyond the 30 s scoring guidelines requires a drastic shift in methodology apart from a change in mindset. The rise of machine learning, and in particular deep learning during the last decade, has prompted new and powerful ways to learn from data, that could be useful for developing future intelligent sleep analysis algorithms. Coupled with the raw amount of data available from a sleep study with relatively few outcome labels, it seems highly probable that unsupervised learning will become increasingly popular for discovering new information about the processes and mechanisms of sleep. Self-supervised representation learning is one such form of unsupervised learning, where a model is incentivized to define latent feature spaces given some data without any associated target labels. The goal of defining such latent representations is to model the underlying data distribution for use in a later downstream task, which has been shown to work effectively for both audio, video and natural language processing domains [216], [217]. Some studies have already been published made regarding the use of such self-supervised techniques for improving sleep stage scoring [218]. The authors were able to effectively model the underlying data distribution of PSG recordings to improve sleep stage scoring in cases where the amount of data was not sufficient to effectively train traditional machine learning or deep learning-based models [218]. As this type of modeling is, in a sense, agnostic to the downstream task at hand, this could be a potential avenue for future research aiming to design "one-shot-analysis"-type models, which is capable of doing everything at once.

A downstream task is a secondary task that a model is not originally trained to complete.

This could e. g. be a model that is able to detect micro-sleep events, predict sleep disorders, recommend treatment plans, etc..

However, it is naïve to think that clinician experience would be redundant in the future of clinical sleep medicine due to automated systems. Semi-supervised learning systems could benefit from having both the robustness and objectivity of the machine, as well as the expertise and flexibility of a trained professional. Recently, *explainable AI* has emerged as a possible tool for further analysis of deep learning models [219], and could also offer new insight into sleep analysis [81].

7 CONCLUSION

Did we learn a lesson here I'm not seeing?

— Summer Sanchez
Rick and Morty, season 1, episode 9

The main focus of this thesis was *to develop a system based on artificial intelligence, that can assist clinicians in the analysis of sleep studies*. The application of such a system would benefit clinicians and patients by shifting time spent on analysis towards patient care. The system was realized in three parts, which each addresses a specific sub-hypothesis:

RH 1: *Advanced biomedical signal processing and machine learning algorithms can be used for efficient, high-performing analysis of sleep studies with regards to sleep stages.*

Chapter 3 presented the MASSC model for automatic classification of sleep stages using raw EEG, EOG, and chin EMG data extracted from PSG. This model was initially proposed in [72], where it was trained and tested on 1850 and 230 PSGs, respectively, yielding an accuracy of 84%. Increasing both the volume and diversity of the data increased performance to 87%, which was described in Section 3.3.

Also presented was the STAGES model for sleep stage classification based on cross-correlation representations of the EEG, EOG and chin EMG and an ensemble of deep neural networks. This model was trained on 2784 PSGs and subsequently validated against six technicians on 70 PSGs, where it outperformed all technicians both on a biased and unbiased consensus score. The model was furthermore found stable with respect to underlying sleep pathologies in all cases except for narcolepsy, which was exploited in Chapter 5.

These findings suggest that AI-based systems such as deep neural networks can augment clinicians with sleep stage classification, and that volume and diversity in datasets are key to high performance.

RH 2: *Advanced biomedical signal processing and machine learning algorithms can be used for efficient, high-performing analysis of sleep studies with regards to sleep events.*

Chapter 4 presented the MSED model for sleep event detection. An initial version of the model was described in Section 4.2 for detection of arousals and leg movements, which was augmented to include sleep disordered breathing events in Section 4.3, using raw EEG, EOG, chin and leg EMG, and respiratory data extracted from the PSG. Training and testing the model on 1653 and 1000 PSGs, respectively, yielded F1 scores of 0.70, 0.63, and 0.62 for arousal, leg movement, and sleep disordered breathing event detection, respectively. The performance was higher when detecting events jointly compared to corresponding single-event models. Index values computed from detected events correlated well with manual annotations with r^2-values of 0.73, 0.77, and 0.78, respectively.

The MSED model was applied in a transfer learning setup under the channel mismatch problem, where the target dataset contained only one EEG channel. This problem has been investigated previously for the case of sleep stage classification, but remains under-investigated in sleep event detection. Using a network pre-trained on a full montage of input channels, stripped of the input processing layers, and subsequently fine-tuned on a smaller dataset allowed for recovery of F1 performance compared to the full montage model.

RH 3: *Advanced biomedical signal processing and machine learning algorithms can be used for efficient, high-performing analysis of sleep studies with regards to sleep disorders*. Chapter 5 presented the application of the STAGES for narcolepsy detection. The motivation behind this was the finding in Section 3.4, which led to the development of a probabilistic model using features derived from the hypnogram representation of a PSG and a Gaussian process classification algorithm. The STAGES model was able to classify NT1 patients with a 91% sensitivity and 96% specificity, which was increased to 99% by adding HLA-DQB1*06:02 typing. This was replicated in an independent cohort of data from two continents for an optimized sensitivity and specificity of 94% and 94%. These findings match the current gold standard for narcolepsy diagnosis while having the benefit of using the PSG alone or in combination with blood samples.

Altogether, the findings presented in this thesis underline the practicality and utility of incorporating AI-based systems in the sleep clinic by providing fast and accurate analysis of sleep stages, sleep events, and sleep disorders.

8 FUTURE WORK

Think for yourselves, don't be sheep.

— Rick Sanchez
Rick and Morty, season 2, episode 11

This thesis has proposed novel methods for sleep stage classification, sleep event detection, and identification of patients with narcolepsy, which may aid clinicians in their work. However, the development of automatic, AI-based systems for clinical sleep analysis is a fast-growing, open-ended field with room for further investigations. An incomplete list of suggestions for future directions of research within computational sleep science is provided here:

- Although the methods described in Chapters 3 to 5 could be described as a single *system*, future research should investigate the development of a unified model, that can make predictions on a small segment scale (sleep stages, micro-events) and a larger, whole-PSG scale (outcome modeling, identification of sleep disorders).
- Unsupervised and semi-supervised learning approaches such as contrastive predicting coding should be investigated with the aim of creating a flexible *generic* model capable of completing multiple several downstream tasks.
- The transfer learning experiments in Section 4.4 was only tested within one cohort; the findings should be replicated across multiple datasets.
- The MSED utility wrt. sleep disorder characterization, classification and prediction should be investigated more thoroughly.
- Although the methods described in this thesis have been tested in data with good performance, these results may still be biased towards the applied cohorts. To alleviate this, the sleep science community should consider establishing publicly available benchmark datasets for independent validation of algorithms.
- The narcolepsy detector presented in Chapter 5 should be validated in more samples from multiple international sleep labs.

- In developing the narcolepsy detector, we did not investigate the inclusion of MSLT findings, due to the specific scope. However, it would be valuable to investigate the relationship between objective findings such as MSLT and the algorithm output.

- It should also be investigated how stable the narcolepsy detection is with regards to the first night effect, repeated PSG recordings, and/or if the inclusion of multiple nights can add value.

- It should be investigated whether the framework for narcolepsy detection could be adapted for detection of other hypersomnias, either in a separate system or as a detection algorithm for central hypersomnias in general.

BIBLIOGRAPHY

[1] American Academy of Sleep Medicine, *International classification of sleep disorders*, 3rd ed. Darien, Il: American Academy of Sleep Medicine, 2014.

[2] T. Young, P. E. Peppard, and D. J. Gottlieb, "Epidemiology of Obstructive Sleep Apnea: A Population Health Perspective", *Am. J. Respir. Crit. Care Med.*, vol. 165, no. 9, pp. 1217–1239, 2002. DOI: 10.1164/rccm.2109080.

[3] D. J. Gottlieb and N. M. Punjabi, "Diagnosis and Management of Obstructive Sleep Apnea", *JAMA*, vol. 323, no. 14, pp. 1389–1400, 2020. DOI: 10.1001/jama.2020.3514.

[4] S. Tufik, R. Santos-Silva, J. A. Taddei, and L. R. A. Bittencourt, "Obstructive Sleep Apnea Syndrome in the Sao Paulo Epidemiologic Sleep Study", *Sleep Med.*, vol. 11, no. 5, pp. 441–446, 2010. DOI: 10.1016/j.sleep.2009.10.005.

[5] R. Heinzer, S. Vat, P. Marques-Vidal, H. Marti-Soler, D. Andries, N. Tobback, V. Mooser, M. Preisig, A. Malhotra, G. Waeber, P. Vollenweider, M. Tafti, and J. Haba-Rubio, "Prevalence of sleep-disordered breathing in the general population: the HypnoLaus study", *Lancet Respir. Med.*, vol. 3, no. 4, 2015. DOI: 10.1016/S2213-2600(15)00043-0.

[6] E. S. Arnardottir, E. Bjornsdottir, K. A. Olafsdottir, B. Benediktsdottir, and T. Gislason, "Obstructive sleep apnoea in the general population: highly prevalent but minimal symptoms", *Eur. Respir. J.*, vol. 47, no. 1, pp. 194–202, 2016. DOI: 10.1183/13993003.01148-2015.

[7] I. Fietze, N. Laharnar, A. Obst, R. Ewert, S. B. Felix, C. Garcia, S. Gläser, M. Glos, C. O. Schmidt, B. Stubbe, H. Völzke, S. Zimmermann, and T. Penzel, "Prevalence and association analysis of obstructive sleep apnea with gender and age differences - Results of SHIP-Trend", *J. Sleep Res.*, vol. 28, no. 5, e12770, 2019. DOI: 10.1111/jsr.12770.

[8] R. J. Ozminkowski, S. Wang, and J. K. Walsh, "The Direct and Indirect Costs of Untreated Insomnia in Adults in the United States", *Sleep*, vol. 30, no. 3, pp. 263–273, 2007. DOI: 10.1093/sleep/30.3.263.

[9] M. M. Ohayon, "Epidemiology of insomnia: what we know and what we still need to learn", *Sleep Med. Rev.*, vol. 6, no. 2, pp. 97–111, 2002. DOI: 10.1053/smrv.2002.0186.

[10] M. J. Decker, J.-M. S. Lin, H. Tabassum, and W. C. Reeves, "Hypersomnolence and Sleep-related Complaints in Metropolitan, Urban, and Rural Georgia", *Am. J. Epidemiol.*, vol. 169, no. 4, pp. 435–443, 2008. DOI: 10.1093/aje/kwn365.

[11] M. M. Ohayon, "Epidemiological Overview of sleep Disorders in the General Population", *Sleep Med. Res.*, vol. 2, no. 1, pp. 1–9, 2011. DOI: 10.17241/smr.2011.2.1.1.

[12] D. A. Johnson, N. Guo, M. Rueschman, R. Wang, J. G. Wilson, and S. Redline, "Prevalence and correlates of obstructive sleep apnea among African Americans: the Jackson Heart Sleep Study", *Sleep*, vol. 41, no. 10, pp. 1–9, 2018. DOI: 10.1093/sleep/zsy154.

[13] N. M. Punjabi, “The Epidemiology of Adult Obstructive Sleep Apnea”, *Proc. Am. Thorac. Soc.*, vol. 5, no. 2, pp. 136–143, 2008. DOI: 10.1513/pats.200709-155MG.

[14] J. M. Lyznicki, T. C. Doege, R. M. David, M. A. Williams, and A. M. A. Council on Scientific Affairs, “Sleepiness, Driving, and Motor Vehicle Crashes”, *JAMA*, vol. 279, no. 23, pp. 1908–1913, 1998. DOI: 10.1001/jama.279.23.1908.

[15] L. J. Findley, M. E. Unverzagt, and P. M. Suratt, “Automobile Accidents Involving Patients with Obstructive Sleep Apnea”, *Am. Rev. Respir. Dis.*, vol. 138, no. 2, pp. 337–340, 1988. DOI: 10.1164/ajrccm/138.2.337.

[16] D. R. Hillman, A. S. Murphy, R. Antic, and L. Pezzullo, “The Economic Cost of Sleep Disorders”, *Sleep*, vol. 29, no. 3, pp. 299–305, 2006. DOI: 10.1093/sleep/29.3.299.

[17] M. Hafner, M. Stepanek, J. Taylor, W. M. Troxel, and C. van Stolk, “Why Sleep Matters-The Economic Costs of Insufficient Sleep: A Cross-Country Comparative Analysis.”, *Rand Heal. Q.*, vol. 6, no. 4, p. 11, 2017.

[18] J. P. Kiley, M. J. Twery, and G. H. Gibbons, “The National Center on Sleep Disorders Research—progress and promise”, *Sleep*, vol. 42, no. 6, pp. 1–5, 2019. DOI: 10.1093/sleep/zsz105.

[19] H. W. Agnew, W. B. Webb, and R. L. Williams, “The first night effect: an EEG study of sleep”, *Psychophysiology*, vol. 2, no. 3, pp. 263–266, 1966. DOI: 10.1111/j.1469-8986.1966.tb02650.x.

[20] S. Scholle, H.-C. Scholle, A. Kemper, S. Glaser, B. Rieger, G. Kemper, and G. Zwacka, “First night effect in children and adolescents undergoing polysomnography for sleep-disordered breathing”, *Clin. Neurophysiol.*, vol. 114, no. 11, pp. 2138–2145, 2003. DOI: 10.1016/S1388-2457(03)00209-8.

[21] M. J. Drinnan, A. Murray, C. J. Griffiths, and G. J. Gibson, “Interobserver Variability in Recognizing Arousal in Respiratory Sleep Disorders”, *Am. J. Respir. Crit. Care Med.*, vol. 158, pp. 358–362, 1998. DOI: 10.1164/ajrccm.158.2.9705035.

[22] C. W. Whitney, D. J. Gottlieb, S. Redline, R. G. Norman, R. R. Dodge, E. Shahar, S. Surovec, and F. J. Nieto, “Reliability of scoring respiratory disturbance indices and sleep staging”, *Sleep*, vol. 21, no. 7, pp. 749–757, 1998. DOI: 10.1093/sleep/21.7.749.

[23] H. Danker-Hopfe, D. Kunz, G. Gruber, G. Klösch, J. L. Lorenzo, S. L. Himanen, B. Kemp, T. Penzel, J. Röschke, H. Dorn, A. Schlögl, E. Trenker, and G. Dorffner, “Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders”, *J. Sleep Res.*, vol. 13, pp. 63–69, 2004. DOI: 10.1046/j.1365-2869.2003.00375.x.

[24] R. G. Norman, I. Pal, C. Stewart, J. A. Walsleben, and D. M. Rapoport, “Interobserver Agreement Among Sleep Scorers From Different Centers in a Large Dataset”, *Sleep*, vol. 23, no. 7, pp. 1–8, 2000. DOI: 10.1093/sleep/23.7.1e.

[25] 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*, vol. 36, no. 4, pp. 591–596, 2013. DOI: 10.5665/sleep.2552.

[26] R. S. Rosenberg and S. Van Hout, "The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring", *J. Clin. Sleep Med.*, vol. 9, pp. 81–87, 2013. DOI: 10.5664/jcsm.2350.

[27] ——, "The American Academy of Sleep Medicine Inter-scorer Reliability Program: Respiratory Events", *J. Clin. Sleep Med.*, vol. 10, no. 4, pp. 447–454, 2014. DOI: 10.5664/jcsm.3630.

[28] M. Younes, J. Raneri, and P. Hanly, "Staging sleep in polysomnograms: Analysis of inter-scorer variability", *J. Clin. Sleep Med.*, vol. 12, no. 6, pp. 885–894, 2016. DOI: 10.5664/jcsm.5894.

[29] M. Younes, S. T. Kuna, A. I. Pack, J. K. Walsh, C. A. Kushida, B. Staley, and G. W. Pien, "Reliability of the American Academy of Sleep Medicine Rules for Assessing Sleep Depth in Clinical Practice", *J. Clin. Sleep Med.*, vol. 14, no. 2, pp. 205–213, 2018. DOI: 10.5664/jcsm.6934.

[30] I. Perez-Pozuelo, B. Zhai, J. Palotti, R. Mall, M. Aupetit, J. M. Garcia-Gomez, S. Taheri, Y. Guan, and L. Fernandez-Luque, "The future of sleep health: a data-driven revolution in sleep science and medicine", *npj Digit. Med.*, vol. 3, p. 42, 2020. DOI: 10.1038/s41746-020-0244-4.

[31] C. M. Depner, P. C. Cheng, J. K. Devine, S. Khosla, M. de Zambotti, R. Robillard, A. Vakulin, and S. P. A. Drummond, "Wearable Technologies for Developing Sleep and Circadian Biomarkers: A Summary of Workshop Discussions", *Sleep*, vol. 43, no. 2, pp. 1–13, 2019. DOI: 10.1093/sleep/zsz254.

[32] B. Şen, M. Peker, A. Çavuşoğlu, and F. V. Çelebi, "A Comparative Study on Classification of Sleep Stage Based on EEG Signals Using Feature Selection and Classification Algorithms", *J. Med. Syst.*, vol. 38, no. 3, p. 18, 2014. DOI: 10.1007/s10916-014-0018-0.

[33] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *Nature*, vol. 521, pp. 436–444, 2015. DOI: 10.1038/nature14539.

[34] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare", *Nat. Biomed. Eng.*, vol. 2, no. 10, pp. 719–731, 2018. DOI: 10.1038/s41551-018-0305-z.

[35] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, "The practical implementation of artificial intelligence technologies in medicine", *Nat. Med.*, vol. 25, no. 1, pp. 30–36, 2019. DOI: 10.1038/s41591-018-0307-0.

[36] M. Ronzhina, O. Janoušek, J. Kolářová, M. Nováková, P. Honzík, and I. Provazník, "Sleep scoring using artificial neural networks", *Sleep Med. Rev.*, vol. 16, no. 3, pp. 251–263, 2012. DOI: 10.1016/j.smrv.2011.06.003.

[37] M. Radha, G. Garcia-Molina, M. Poel, and G. Tononi, "Comparison of feature and classifier algorithms for online automatic sleep staging based on a single EEG signal", *2014 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Chicago, IL, USA: IEEE, 2014, pp. 1876–1880. DOI: 10.1109/EMBC.2014.6943976.

[38] K. Aboalayon, M. Faezipour, W. Almuhammadi, and S. Moslehpour, "Sleep Stage Classification Using EEG Signal Analysis: A Comprehensive Survey and New Investigation", *Entropy*, vol. 18, no. 9, p. 272, 2016. DOI: 10.3390/e18090272.

[39] R. Boostani, F. Karimzadeh, and M. Nami, "A comparative review on sleep stage classification methods in patients and healthy individuals", *Comput. Methods Programs Biomed.*, vol. 140, pp. 77–91, 2017. DOI: 10.1016/j.cmpb.2016.12.004.

[40] R. E. Brown, R. Basheer, J. T. McKenna, R. E. Strecker, and R. W. McCarley, "Control of Sleep and Wakefulness", *Physiol. Rev.*, vol. 92, no. 3, pp. 1087–1187, 2012. DOI: 10.1152/physrev.00032.2011.

[41] C. B. Saper, P. M. Fuller, N. P. Pedersen, J. Lu, and T. E. Scammell, "Sleep State Switching", *Neuron*, vol. 68, no. 6, pp. 1023–1042, 2010. DOI: 10.1016/j.neuron.2010.11.032.

[42] M. A. Carskadon and W. C. Dement, "Normal Human Sleep: An Overview", *Princ. Pract. Sleep Med.* M. H. Kryger, T. Roth, and W. C. Dement, Eds., 5th ed., St. Louis, MO, USA: Elsevier Inc., 2011, pp. 16–26.

[43] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. L. Marcus, and B. V. Vaughn, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.6*. Darien, IL, USA: American Academy of Sleep Medicine, 2020, **AASM2020**.

[44] M. Cesari, "Data-driven classification algorithms for identification and characterization of early neurodegeneration", PhD, Technical University of Denmark, 2019, p. 182.

[45] P. Simor, G. van der Wijk, L. Nobili, and P. Peigneux, "The microstructure of REM sleep: why phasic and tonic?", *Sleep Med. Rev.*, 2020. DOI: 10.1016/j.smrv.2020.101305.

[46] W. D. Foulkes, "Dream reports from different stages of sleep.", *J. Abnorm. Soc. Psychol.*, vol. 65, no. 1, pp. 14–25, 1962. DOI: 10.1037/h0040431.

[47] J. A. Hobson, "REM sleep and dreaming: towards a theory of protoconsciousness", *Nat. Rev. Neurosci.*, vol. 10, no. 11, pp. 803–813, 2009. DOI: 10.1038/nrn2716.

[48] P. McNamara, P. Johnson, D. McLaren, E. Harris, C. Beauharnais, and S. Auerbach, "REM And NREM Sleep Mentation", *Int. Rev. Neurobiol.*, vol. 92, pp. 69–86, 2010. DOI: 10.1016/S0074-7742(10)92004-7.

[49] F. Siclari, B. Baird, L. Perogamvros, G. Bernardi, J. J. LaRocque, B. Riedner, M. Boly, B. R. Postle, and G. Tononi, "The neural correlates of dreaming", *Nat. Neurosci.*, vol. 20, no. 6, pp. 872–878, 2017. DOI: 10.1038/nn.4545.

[50] M. Takahara, S. Kanayama, and T. Hori, "Co-occurrence of Sawtooth Waves and Rapid Eye Movements during REM Sleep", *Int. J. Bioelectromagn.*, vol. 11, no. 3, pp. 144–148, 2009.

[51] U. Ermis, K. Krakow, and U. Voss, "Arousal thresholds during human tonic and phasic REM sleep", *J. Sleep Res.*, vol. 19, no. 3, pp. 400–406, 2010. DOI: 10.1111/j.1365-2869.2010.00831.x.

[52] M. Sallinen, J. Kaartinen, and H. Lyytinen, "Processing of auditory stimuli during tonic and phasic periods of REM sleep as revealed by event-related brain potentials", *J. Sleep Res.*, vol. 5, pp. 220–228, 1996. DOI: 10.1111/j.1365-2869.1996.00220.x.

[53] M. Takahara, H. Nittono, and T. Hori, "Comparison of the event-related potentials between tonic and phasic periods of rapid eye movement sleep", *Psychiatry Clin. Neurosci.*, vol. 56, pp. 257–258, 2002. DOI: 10.1046/j.1440-1819.2002.00999.x.

[54] ——, "Effect of Voluntary Attention on Auditory Processing During REM Sleep", *Sleep*, vol. 29, no. 7, pp. 975–982, 2006. DOI: 10.1093/sleep/29.7.975.

[55] R. Manni, M. Terzaghi, and M. Glorioso, "Motor-Behavioral Episodes in REM Sleep Behavior Disorder and Phasic Events During REM Sleep", *Sleep*, vol. 32, no. 2, pp. 241–245, 2009. DOI: 10.1093/sleep/32.2.241.

[56] F. De Carli, P. Proserpio, E. Morrone, I. Sartori, M. Ferrara, S. A. Gibbs, L. De Gennaro, G. Lo Russo, and L. Nobili, "Activation of the motor cortex during phasic rapid eye movement sleep", *Ann. Neurol.*, vol. 79, pp. 326–330, 2016. DOI: 10.1002/ana.24556.

[57] J.-S. Sunwoo, K. S. Cha, J.-I. Byun, T.-J. Kim, J.-S. Jun, J.-A. Lim, S.-T. Lee, K.-H. Jung, K.-I. Park, K. Chu, H.-J. Kim, M. Kim, S. K. Lee, K. H. Kim, C. H. Schenck, and K.-Y. Jung, "Abnormal activation of motor cortical network during phasic REM sleep in idiopathic REM sleep behavior disorder", *Sleep*, vol. 42, no. 2, pp. 1–10, 2019. DOI: 10.1093/sleep/zsy227.

[58] L. Schneider, "Anatomy and Physiology of Normal Sleep", *Sleep Neurol. Dis.* San Diego, CA, USA: Academic Press, 2017, pp. 1–28. DOI: 10.1016/B978-0-12-804074-4.00001-7.

[59] C. Peyron, D. K. Tighe, A. N. van den Pol, L. de Lecea, H. C. Heller, J. G. Sutcliffe, and T. S. Kilduff, "Neurons Containing Hypocretin (Orexin) Project to Multiple Neuronal Systems", *J. Neurosci.*, vol. 18, no. 23, pp. 9996–10 015, 1998. DOI: 10.1523/JNEUROSCI.18-23-09996.1998.

[60] C. B. Saper, T. C. Chou, and T. E. Scammell, "The sleep switch: hypothalamic control of sleep and wakefulness", *Trends Neurosci.*, vol. 24, no. 12, pp. 726–731, 2001. DOI: 10.1016/S0166-2236(00)02002-6.

[61] C. B. Saper, T. E. Scammell, and J. Lu, "Hypothalamic regulation of sleep and circadian rhythms", *Nature*, vol. 437, no. 7063, pp. 1257–1263, 2005. DOI: 10.1038/nature04284.

[62] 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. DOI: 10.5664/jcsm.2352.

[63] J. Santamaria, B. Höogl, C. Trenkwalder, and D. Bliwise, "Scoring Sleep in Neurological Patients: The Need for Specific Considerations", *Sleep*, vol. 34, no. 10, pp. 1283–1284, 2011. DOI: 10.5665/SLEEP.1256.

[64] B. R. Kornum, S. Knudsen, H. M. Ollila, F. Pizza, P. J. Jennum, Y. Dauvilliers, and S. Overeem, "Narcolepsy", *Nat. Rev. Dis. Prim.*, vol. 3, no. 1, p. 16 100, 2017. DOI: 10.1038/nrdp.2016.100.

[65] K. B. Mikkelsen, D. B. Villadsen, M. Otto, and P. Kidmose, "Automatic sleep staging using ear-EEG", *Biomed. Eng. Online*, vol. 16, p. 111, 2017. DOI: 10.1186/s12938-017-0400-5.

[66] K. B. Mikkelsen, Y. R. Tabar, S. L. Kappel, C. B. Christensen, H. O. Toft, M. C. Hemmsen, M. L. Rank, M. Otto, and P. Kidmose, "Accurate whole-night sleep monitoring with dry-contact ear-EEG", *Sci. Rep.*, vol. 9, p. 16 824, 2019. DOI: 10.1038/s41598-019-53115-3.

[67] 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 Breath.*, vol. 19, no. 1, pp. 191–195, 2015. DOI: 10.1007/s11325-014-0990-0.

[68] M. H. Bonnet, K. Doghramji, T. Roehrs, E. J. Stepanski, S. H. Sheldon, A. S. Walters, M. Wise, and A. L. Chesson, "The scoring of arousal in sleep: Reliability, validity, and alternatives", *J. Clin. Sleep Med.*, vol. 3, no. 2, pp. 133–145, 2007. DOI: 10.5664/jcsm.26815.

[69] J. S. Loredo, J. L. Clausen, S. Ancoli-Israel, and J. E. Dimsdale, "Night-to-Night Arousal Variability and Interscorer Reliability of Arousal Measurements", *Sleep*, vol. 22, no. 7, pp. 916–920, 1999. DOI: 10.1093/sleep/22.7.916.

[70] M. Smurra, M. Dury, G. Aubert, D. Rodenstein, and G. Liistro, "Sleep fragmentation: comparison of two definitions of short arousals during sleep in OSAS patients", *Eur. Respir. J.*, vol. 17, pp. 723–727, 2001. DOI: 10.1183/09031936.01.17407230.

[71] R. J. Thomas, "Arousals in Sleep-disordered Breathing: Patterns and Implications", *Sleep*, vol. 26, no. 8, pp. 1042–1047, 2003. DOI: 10.1093/sleep/26.8.1042.

[72] **A. N. Olesen**, P. Jennum, P. Peppard, E. Mignot, and H. B. D. Sorensen, "Deep residual networks for automatic sleep stage classification of raw polysomnographic waveforms", *2018 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Honolulu, HI, USA: IEEE, 2018. DOI: 10.1109/EMBC.2018.8513080.

[73] J. B. Stephansen*, **A. N. Olesen***, M. Olsen, A. Ambati, E. B. Leary, H. E. Moore, O. Carrillo, L. Lin, F. Han, H. Yan, Y. L. Sun, Y. Dauvilliers, S. Scholz, L. Barateau, B. Hogl, A. Stefani, S. C. Hong, T. W. Kim, F. Pizza, G. Plazzi, S. Vandi, E. Antelmi, D. Perrin, S. T. Kuna, P. K. Schweitzer, C. Kushida, P. E. Peppard, H. B. D. Sorensen, P. Jennum, and E. Mignot, "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy", *Nat. Commun.*, vol. 9, p. 5229, 2018. DOI: 10.1038/s41467-018-07229-3.

[74] **A. N. Olesen**, P. Jennum, E. Mignot, and H. B. D. Sorensen, *Automatic sleep stage classification with deep residual networks in a mixed-cohort setting*, 2020, (*under review*).

[75] W. Chiao and M. L. Durr, "Trends in sleep studies performed for Medicare beneficiaries", *Laryngoscope*, vol. 127, no. 12, pp. 2891–2896, 2017. DOI: 10.1002/lary.26736.

[76] 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*. Darien, IL, USA: American Academy of Sleep Medicine, 2018, **AASM**2018.

[77] M. Younes, “The case for using digital EEG analysis in clinical sleep medicine”, *Sleep Sci. Pract.*, vol. 1, no. 1, p. 2, 2017. DOI: 10.1186/s41606-016-0005-0.

[78] L. Fiorillo, A. Puiatti, M. Papandrea, P.-L. Ratti, P. Favaro, C. Roth, P. Bargiotas, C. L. Bassetti, and F. D. Faraci, “Automated sleep scoring: A review of the latest approaches”, *Sleep Med. Rev.*, vol. 48, p. 101 204, 2019. DOI: 10.1016/j.smrv.2019.07.007.

[79] 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”, *Circulation*, vol. 101, no. 23, e215–e220, 2000. DOI: 10.1161/01.CIR.101.23.e215.

[80] B. Kemp, A. Zwinderman, B. Tuk, H. Kamphuisen, and J. Oberye, “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, 2000. DOI: 10.1109/10.867928.

[81] A. Vilamala, K. H. Madsen, and L. K. Hansen, “Deep convolutional neural networks for interpretable analysis of EEG sleep stage scoring”, *2017 IEEE 27th Int. Work. Mach. Learn. Signal Process.*, Tokyo, Japan: IEEE, 2017, pp. 1–6. DOI: 10.1109/MLSP.2017.8168133. arXiv: 1710.00633 [cs.CV].

[82] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. D. Vos, “Automatic Sleep Stage Classification Using Single-Channel EEG: Learning Sequential Features with Attention-Based Recurrent Neural Networks”, *2018 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Honolulu, HI, USA: IEEE, 2018, pp. 1452–1455. DOI: 10.1109/EMBC.2018.8512480.

[83] 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, 2017. DOI: 10.1109/TNSRE.2017.2721116.

[84] 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”, *J. Sleep Res.*, vol. 23, no. 6, pp. 628–635, 2014. DOI: 10.1111/jsr.12169.

[85] 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, 2018. DOI: 10.1109/TNSRE.2018.2813138.

[86] 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”, *2018 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Honolulu, HI, USA: IEEE, 2018, pp. 171–174. DOI: 10.1109/EMBC.2018.8512214.

[87] 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, 2019. DOI: 10.1109/TBME.2018.2872652.

[88] ——, “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, 2019. DOI: 10.1109/TNSRE.2019.2896659.

[89] S. Biswal, H. Sun, B. Goparaju, M. B. Westover, J. Sun, and M. T. Bianchi, “Expert-level sleep scoring with deep neural networks”, *J. Am. Med. Informatics Assoc.*, vol. 25, no. 12, pp. 1643–1650, 2018. DOI: 10.1093/jamia/ocy131.

[90] 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*, vol. 41, no. 5, pp. 1–11, 2018. DOI: 10.1093/sleep/zsy041.

[91] S. Biswal, J. Kulas, H. Sun, B. Goparaju, M. B. Westover, M. T. Bianchi, and J. Sun, “SLEEPNET: Automated Sleep Staging System via Deep Learning”, 2017. arXiv: 1707.08262 [cs.LG].

[92] T. Young, M. Palta, J. Dempsey, J. Skatrud, S. Weber, and S. Badr, “The Occurrence of Sleep-Disordered Breathing among Middle-Aged Adults”, *N. Engl. J. Med.*, vol. 328, no. 17, pp. 1230–1235, 1993. DOI: 10.1056/NEJM199304293281704.

[93] T. Young, L. Finn, P. E. Peppard, M. Szklo-Coxe, D. Austin, J. Nieto, R. Stubss, and K. M. Hla, “Sleep Disordered Breathing and Mortality: Eighteen-Year Follow-up of the Wisconsin Sleep Cohort”, *Sleep*, vol. 31, no. 8, pp. 291–292, 2008. DOI: 10.5665/sleep/31.8.1071.

[94] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, R. M. Lloyd, C. L. Marcus, and B. V. Vaughn, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.3*. Darien, IL, USA: The American Academy of Sleep Medicine, 2016, **AASM2016**.

[95] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition”, *Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[96] ——, “Identity Mappings in Deep Residual Networks”, *Comput. Vis. – ECCV 2016*, 2016, pp. 630–645. DOI: 10.1007/978-3-319-46493-0_38.

[97] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization”, 2014. arXiv: 1412.6980 [cs.LG].

[98] K. He, X. Zhang, S. Ren, and J. Sun, “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”, *Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034. DOI: 10.1109/ICCV.2015.123.

[99] H. Koch, J. A. Christensen, R. Frandsen, M. Zoetmulder, L. Arvastson, S. R. Christensen, P. Jennum, and H. B. Sorensen, “Automatic sleep classification using a data-driven topic model reveals latent sleep states”, *J. Neurosci. Methods*, vol. 235, pp. 130–137, 2014. DOI: 10.1016/j.jneumeth.2014.07.002.

[100] J. A. Christensen, M. Zoetmulder, H. Koch, R. Frandsen, L. Arvastson, S. R. Christensen, P. Jennum, and H. B. Sorensen, “Data-driven modeling of sleep EEG and EOG reveals characteristics indicative of pre-Parkinson’s and Parkinson’s disease”, *J. Neurosci. Methods*, vol. 235, pp. 262–276, 2014. DOI: 10.1016/j.jneumeth.2014.07.014.

[101] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision", *2016 IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA: IEEE, 2016, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308.

[102] S. Khalighi, T. Sousa, J. M. Santos, and U. Nunes, "ISRUC-Sleep: A comprehensive public dataset for sleep researchers", *Comput. Methods Programs Biomed.*, vol. 124, pp. 180–192, 2016. DOI: 10.1016/j.cmpb.2015.10.013.

[103] J. B. Blank, P. M. Cawthon, M. L. Carrion-Petersen, L. Harper, J. P. Johnson, E. Mitson, and R. R. Delay, "Overview of recruitment for the osteoporotic fractures in men study (MrOS)", *Contemp. Clin. Trials*, vol. 26, no. 5, pp. 557–568, 2005. DOI: 10.1016/j.cct.2005.05.005.

[104] E. Orwoll, J. B. Blank, E. Barrett-Connor, J. Cauley, S. Cummings, K. Ensrud, C. Lewis, P. M. Cawthon, R. Marcus, L. M. Marshall, J. McGowan, K. Phipps, S. Sherman, M. L. Stefanick, and K. Stone, "Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study — A large observational study of the determinants of fracture in older men", *Contemp. Clin. Trials*, vol. 26, no. 5, pp. 569–585, 2005. DOI: 10.1016/j.cct.2005.05.006.

[105] T. Blackwell, K. Yaffe, S. Ancoli-Israel, S. Redline, K. E. Ensrud, M. L. Stefanick, A. Laffan, and K. L. Stone, "Associations Between Sleep Architecture and Sleep-Disordered Breathing and Cognition in Older Community-Dwelling Men: The Osteoporotic Fractures in Men Sleep Study", *J. Am. Geriatr. Soc.*, vol. 59, no. 12, pp. 2217–2225, 2011. DOI: 10.1111/j.1532-5415.2011.03731.x.

[106] A. Rechtschaffen and A. Kales, Eds., *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. Washington, DC: National Institute of Health, 1968.

[107] D. A. Dean, A. L. Goldberger, R. Mueller, M. Kim, M. Rueschman, D. Mobley, S. S. Sahoo, C. P. Jayapandian, L. Cui, M. G. Morrical, S. Surovec, G.-Q. Zhang, and S. Redline, "Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource", *Sleep*, vol. 39, no. 5, pp. 1151–1164, 2016. DOI: 10.5665/sleep.5774.

[108] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline, "The National Sleep Research Resource: towards a sleep data commons", *J. Am. Med. Informatics Assoc.*, vol. 25, no. 10, pp. 1351–1358, 2018. DOI: 10.1093/jamia/ocy064.

[109] S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet, P. W. Wahl, V. Howard, C. Iber, J. P. Kiley, J. Nieto, G. T. O. Connor, D. M. Rapoport, S. Redline, I. M. Samet, and P. W. Wahl, "The Sleep Heart Health Study: Design, Rationale, and Methods", *Sleep*, vol. 20, pp. 1077–1085, 1997. DOI: 10.1093/sleep/20.12.1077.

[110] S. Redline, M. H. Sanders, B. K. Lind, S. F. Quan, C. Iber, D. J. Gottlieb, W. H. Bonekat, D. M. Rapoport, P. L. Smith, and J. P. Kiley, "Methods for Obtaining and Analyzing Unattended Polysomnography Data for a Multicenter Study", *Sleep*, vol. 21, no. 7, pp. 759–767, 1998. DOI: 10.1093/sleep/21.7.759.

[111] O. Andlauer, H. Moore, L. Jouhier, C. Drake, P. E. Peppard, F. Han, S.-C. Hong, F. Poli, G. Plazzi, R. O'Hara, E. Haffen, T. Roth, T. Young, and E. Mignot, "Nocturnal Rapid Eye Movement Sleep Latency for Identifying Patients With Narcolepsy/Hypocretin Deficiency", *JAMA Neurol.*, vol. 70, no. 7, p. 891, 2013. DOI: 10.1001/jamaneurol.2013.1589.

[112] H. Moore, E. Leary, S.-Y. Lee, O. Carrillo, R. Stubbs, P. Peppard, T. Young, B. Widrow, and E. Mignot, "Design and Validation of a Periodic Leg Movement Detector", *PLoS One*, vol. 9, no. 12, e114565, 2014. DOI: 10.1371/journal.pone.0114565.

[113] S. Chambon, V. Thorey, P. J. Arnal, E. Mignot, and A. Gramfort, "A Deep Learning Architecture to Detect Events in EEG Signals During Sleep", *2018 IEEE 28th Int. Work. Mach. Learn. Signal Process.*, Aalborg, Denmark: IEEE, 2018, pp. 1–6. DOI: 10.1109/MLSP.2018.8517067.

[114] S. Chambon, V. Thorey, P. Arnal, E. Mignot, and A. Gramfort, "DOSED: A deep learning approach to detect multiple sleep micro-events in EEG signal", *J. Neurosci. Methods*, vol. 321, pp. 64–78, 2019. DOI: 10.1016/j.jneumeth.2019.03.017.

[115] **A. N. Olesen**, S. Chambon, V. Thorey, P. Jennum, E. Mignot, and H. B. D. Sorensen, "Towards a Flexible Deep Learning Method for Automatic Detection of Clinically Relevant Multi-Modal Events in the Polysomnogram", *2019 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Berlin, Germany: IEEE, 2019, pp. 556–561. DOI: 10.1109/EMBC.2019.8856570.

[116] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", *Proc. 32nd Int. Conf. Mach. Learn.*, Lille, France: JMLR, 2015. arXiv: 1502.03167 [cs.LG].

[117] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches", *Proc. SSST-8, Eighth Work. Syntax. Semant. Struct. Stat. Transl.*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 103–111. DOI: 10.3115/v1/W14-4012.

[118] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks", *Proc. Thirteen. Int. Conf. Artif. Intell. Stat.*, vol. 9, Sardinia, Italy: PMLR, 2010, pp. 249–256.

[119] A. Brink-Kjaer, **A. N. Olesen**, P. E. Peppard, K. L. Stone, P. Jennum, E. Mignot, and H. B. D. Sorensen, "Automatic Detection of Cortical Arousals in Sleep and their Contribution to Daytime Sleepiness", 2019. arXiv: 1906.01700 [q-bio.NC].

[120] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data.", *Biometrics*, vol. 33, pp. 159–174, 1977.

[121] A. Sors, S. Bonnet, S. Mirek, L. Vercueil, and J. F. Payen, "A convolutional neural network for sleep stage scoring from raw single-channel EEG", *Biomed. Signal Process. Control*, vol. 42, pp. 107–114, 2018. DOI: 10.1016/j.bspc.2017.12.001.

[122] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F. F. Li, "ImageNet Large Scale Visual Recognition Challenge", *Int. J. Comput. Vis.*, pp. 211–252, 2015. DOI: 10.1007/s11263-015-0816-y.

[123] M. Banko and E. Brill, “Scaling to very very large corpora for natural language disambiguation”, *Proc. 39th Annu. Meet. Assoc. Comput. Linguist. - ACL '01*, Morristown, NJ, USA: Association for Computational Linguistics, 2001, pp. 26–33. DOI: 10.3115/1073012.1073017.

[124] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, “Real-time human pose recognition in parts from single depth images”, *CVPR 2011*, vol. 411, Providence, RI, USA: IEEE, 2011, pp. 1297–1304. DOI: 10.1109/CVPR.2011.5995316.

[125] T. Young, M. Palta, J. Dempsey, P. E. Peppard, F. J. Nieto, and K. M. Hla, “Burden of sleep apnea: rationale, design, and major findings of the Wisconsin Sleep Cohort study”, *WMJ*, vol. 108, no. 5, pp. 246–249, 2009.

[126] S. T. Kuna, R. Benca, C. A. Kushida, J. Walsh, M. Younes, B. Staley, A. Hanlon, A. I. Pack, G. W. Pien, and A. Malhotra, “Agreement in Computer-Assisted Manual Scoring of Polysomnograms across Sleep Centers”, *Sleep*, vol. 36, no. 4, pp. 583–589, 2013. DOI: 10.5665/sleep.2550.

[127] S.-C. Hong, L. Lin, J.-H. Jeong, Y.-K. Shin, J.-H. Han, J.-H. Lee, S.-P. Lee, J. Zhang, M. Einen, and E. Mignot, “A Study of the Diagnostic Utility of HLA Typing, CSF Hypocretin-1 Measurements, and MSLT Testing for the Diagnosis of Narcolepsy in 163 Korean Patients With Unexplained Excessive Daytime Sleepiness”, *Sleep*, vol. 29, no. 11, pp. 1429–1438, 2006. DOI: 10.1093/sleep/29.11.1429.

[128] R. K. Malhotra and A. Y. Avidan, “Sleep Stages and Scoring Technique”, *Atlas Sleep Med.* 2nd ed., Elsevier, 2014, ch. 3, pp. 77–99.

[129] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, R. M. Lloyd, 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 2.4*. Darien, IL, USA: American Academy of Sleep Medicine, 2017, **AASM**2017.

[130] R. Caruana, S. Lawrence, and L. Giles, “Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping”, *Proc. 13th Int. Conf. Neural Inf. Process. Syst.*, Denver, CO, USA: MIT Press, 2000, pp. 381–387.

[131] M. H. Silber, S. Ancoli-Israel, M. H. Bonnet, S. Chokroverty, M. M. Grigg-Damberger, M. Hirshkowitz, S. Kapen, S. a. Keenan, M. H. Kryger, T. Penzel, M. R. Pressman, and C. Iber, “The Visual Scoring of Sleep in Adults”, *J. Clin. Sleep Med.*, vol. 03, no. 02, pp. 121–131, 2007. DOI: 10.5664/jcsm.26814.

[132] B. Hjorth, “EEG analysis based on time domain properties”, *Electroencephalogr. Clin. Neurophysiol.*, vol. 29, no. 3, pp. 306–310, 1970. DOI: 10.1016/0013-4694(70)90143-4.

[133] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory”, *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.

[134] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[135] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York City, New York, USA: Springer, 2006.

[136] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators”, *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989. DOI: 10.1016/0893-6080(89)90020-8.

[137] K. Lenc and A. Vedaldi, “Understanding image representations by measuring their equivariance and equivalence”, *2015 IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA: IEEE, 2015, pp. 991–999. DOI: 10.1109/CVPR.2015.7298701.

[138] F. Amzica and M. Steriade, “Electrophysiological correlates of sleep delta waves”, *Electroencephalogr. Clin. Neurophysiol.*, vol. 107, no. 2, pp. 69–83, 1998. DOI: 10.1016/S0013-4694(98)00051-0.

[139] A. Y. Kaplan, A. A. Fingelkurts, A. A. Fingelkurts, S. V. Borisov, and B. S. Darkhovsky, “Nonstationary nature of the brain activity as revealed by EEG/MEG: Methodological, practical and conceptual challenges”, *Signal Processing*, vol. 85, no. 11, pp. 2190–2212, 2005. DOI: 10.1016/j.sigpro.2005.07.010.

[140] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, 2015. arXiv: 1409.1556 [cs.CV].

[141] B. Polyak, “Some methods of speeding up the convergence of iteration methods”, *USSR Comput. Math. Math. Phys.*, vol. 4, no. 5, pp. 1–17, 1964. DOI: 10.1016/0041-5553(64)90137-5.

[142] A. Krogh and J. A. Hertz, “A Simple Weight Decay Can Improve Generalization”, *Adv. Neural Inf. Process. Syst.*, vol. 4, 1992, pp. 950–957.

[143] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”, *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[144] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups”, *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012. DOI: 10.1109/MSP.2012.2205597.

[145] J.-Z. Cheng, D. Ni, Y.-H. Chou, J. Qin, C.-M. Tiu, Y.-C. Chang, C.-S. Huang, D. Shen, and C.-M. Chen, “Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans”, *Sci. Rep.*, vol. 6, no. 1, p. 24 454, 2016. DOI: 10.1038/srep24454.

[146] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, “Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs”, *JAMA*, vol. 316, no. 22, p. 2402, 2016. DOI: 10.1001/jama.2016.17216.

[147] B. E. Bejnordi, M. Veta, P. J. van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. W. M. van der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, O. Geessink, N. Stathonikos, M. C. van Dijk, P. Bult, F. Beca, A. H. Beck, D. Wang, A. Khosla, R. Gargeya, H. Irshad, A. Zhong, Q. Dou, Q. Li, H. Chen, H.-J. Lin, P.-A. Heng, C. Haß, E. Bruni, Q. Wong, U. Halici, M. Ü. Öner, R. Cetin-Atalay, M. Berseth, V. Khvatkov, A. Vylegzhanin, O. Kraus, M. Shaban, N. Rajpoot, R. Awan, K. Sirinukunwattana, T. Qaiser, Y.-W. Tsang, D. Tellez, J. Annuscheit, P. Hufnagl, M.

Valkonen, K. Kartasalo, L. Latonen, P. Ruusuvuori, K. Liimatainen, S. Albarqouni, B. Mungal, A. George, S. Demirci, N. Navab, S. Watanabe, S. Seno, Y. Takenaka, H. Matsuda, H. Ahmady Phoulady, V. Kovalev, A. Kalinovsky, V. Liauchuk, G. Bueno, M. M. Fernandez-Carrobles, I. Serrano, O. Deniz, D. Racoceanu, and R. Venâncio, "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer", *JAMA*, vol. 318, no. 22, p. 2199, 2017. DOI: 10.1001/jama.2017.14585.

[148] 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*, vol. 542, no. 7639, pp. 115–118, 2017. DOI: 10.1038/nature21056.

[149] P. Lakhani and B. Sundaram, "Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks", *Radiology*, vol. 284, no. 2, pp. 574–582, 2017. DOI: 10.1148/radiol.2017162326.

[150] D. S. W. Ting, C. Y.-L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I. Y. San Yeo, S. Y. Lee, E. Y. M. Wong, C. Sabanayagam, M. Baskaran, F. Ibrahim, N. C. Tan, E. A. Finkelstein, E. L. Lamoureux, I. Y. Wong, N. M. Bressler, S. Sivaprasad, R. Varma, J. B. Jonas, M. G. He, C.-Y. Cheng, G. C. M. Cheung, T. Aung, W. Hsu, M. L. Lee, and T. Y. Wong, "Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes", *JAMA*, vol. 318, no. 22, p. 2211, 2017. DOI: 10.1001/jama.2017.18152.

[151] P. Anderer, G. Gruber, S. Parapatics, M. Woertz, T. Miazhynskaia, 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 × 7 Utilizing the Siesta Database", *Neuropsychobiology*, vol. 51, no. 3, pp. 115–133, 2005. DOI: 10.1159/000085205.

[152] T. Lajnef, S. Chaibi, P. Ruby, P.-E. Aguera, J.-B. Eichenlaub, M. Samet, A. Kachouri, and K. Jerbi, "Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines", *J. Neurosci. Methods*, vol. 250, pp. 94–105, 2015. DOI: 10.1016/j.jneumeth.2015.01.022.

[153] **A. N. Olesen**, J. A. E. Christensen, H. B. D. Sorensen, and P. J. Jennum, "A Noise-Assisted Data Analysis Method for Automatic EOG-Based Sleep Stage Classification Using Ensemble Learning", *38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Orlando, FL, USA: IEEE, 2016, pp. 3769–3772. DOI: 10.1109/EMBC.2016.7591548.

[154] T. L. T. da Silveira, A. J. Kozakevicius, and C. R. Rodrigues, "Single-channel EEG sleep stage classification based on a streamlined set of statistical features in wavelet domain", *Med. Biol. Eng. Comput.*, vol. 55, no. 2, pp. 343–352, 2017. DOI: 10.1007/s11517-016-1519-4.

[155] **A. N. Olesen**, P. Jennum, E. Mignot, and H. B. D. Sorensen, *A multimodal sleep event detection algorithm for clinical sleep analysis*, 2020, (*in preparation*).

[156] **A. N. Olesen**, P. Jennum, E. Mignot, and H. B. D. Sorensen, *Deep transfer learning for improving single-EEG arousal detection*, 2020. arXiv: 2004.05111 [cs.CV], (*accepted*, IEEE EMBC 2020).

[157] **A. N. Olesen***, M. Cesari*, J. A. E. Christensen, H. B. D. Sorensen, E. Mignot, and P. Jennum, "A comparative study of methods for automatic detection of rapid eye movement abnormal muscular activity in narcolepsy", *Sleep Med.*, vol. 44, pp. 97–105, 2018. DOI: 10.1016/j.sleep.2017.11.1141.

[158] A. Brink-Kjaer, **A. N. Olesen**, P. E. Peppard, K. L. Stone, P. Jennum, E. Mignot, and H. B. Sorensen, "Automatic detection of cortical arousals in sleep and their contribution to daytime sleepiness", *Clin. Neurophysiol.*, vol. 131, no. 6, pp. 1187–1203, 2020. DOI: 10.1016/j.clinph.2020.02.027.

[159] L. Carvelli, **A. N. Olesen**, A. Brink-Kjær, E. B. Leary, P. E. Peppard, E. Mignot, H. B. Sørensen, and P. Jennum, "Design of a deep learning model for automatic scoring of periodic and non-periodic leg movements during sleep validated against multiple human experts", *Sleep Medicine*, vol. 69, pp. 109–119, 2020. DOI: 10.1016/j.sleep.2019.12.032.

[160] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", *2016 IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA: IEEE, 2016, pp. 779–788. DOI: 10.1109/CVPR.2016.91. arXiv: 1506.02640 [cs.CV].

[161] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger", 2016. arXiv: 1612.08242 [cs.CV].

[162] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector", *Comput. Vis. – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham, Switzerland: Springer, 2016, pp. 21–37. DOI: 10.1007/978-3-319-46448-0_2.

[163] H. Phan, O. Y. Chén, P. Koch, Z. Lu, I. McLoughlin, A. Mertins, and M. De Vos, "Towards More Accurate Automatic Sleep Staging via Deep Transfer Learning", pp. 1–11, 2019. arXiv: 1907.13177 [cs.LG].

[164] H. Phan, O. Y. Chen, P. Koch, A. Mertins, and M. D. Vos, "Deep Transfer Learning for Single-Channel Automatic Sleep Staging with Channel Mismatch", *2019 27th Eur. Signal Process. Conf.*, A Coruña, Spain: IEEE, 2019, pp. 1–5. DOI: 10.23919/EUSIPCO.2019.8902977.

[165] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch", *31st Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017.

[166] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library", *Adv. Neural Inf. Process. Syst. 32*, 2019, pp. 8024–8035. arXiv: 1912.01703.

[167] 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", *J. Sleep Res.*, vol. 18, no. 1, pp. 74–84, 2009. DOI: 10.1111/j.1365-2869.2008.00700.x.

[168] D. Alvarez-Estevez and I. Fernández-Varela, "Large-scale validation of an automatic EEG arousal detection algorithm using different heterogeneous databases", *Sleep Med.*, vol. 57, pp. 6–14, 2019. DOI: 10.1016/j.sleep.2019.01.025.

[169] American Sleep Disorders Association, "EEG arousals: scoring rules and examples: a preliminary report from the Sleep Disorders Atlas Task Force of the American Sleep Disorders Association", *Sleep*, vol. 15, no. 2, pp. 173–184, 1992, PMID: 11032543.

[170] M. Zucconi, R. Ferri, R. Allen, P. C. Baier, O. Bruni, S. Chokroverty, L. Ferini-Strambi, S. Fulda, D. Garcia-Borreguero, W. A. Hening, M. Hirshkowitz, B. Högl, M. Hornyak, M. King, P. Montagna, L. Parrino, G. Plazzi, and M. G. Terzano, "The official World Association of Sleep Medicine (WASM) standards for recording and scoring periodic leg movements in sleep (PLMS) and wakefulness (PLMW) developed in collaboration with a task force from the International Restless Legs Syndrome Study Grou", *Sleep Med.*, vol. 7, no. 2, pp. 175–183, 2006. DOI: 10.1016/j.sleep.2006.01.001.

[171] Y. A. LeCun, L. Bottou, G. B. Orr, and K. R. Müller, "Efficient backprop", *Neural Networks: Tricks of the Trade*, ser. Lect. Notes Comput. Sci. vol 7700, G. Montavon, G. B. Orr, and K.-R. Müller, Eds., Springer Berlin Heidelberg, 2012. DOI: 10.1007/978-3-642-35289-8-3.

[172] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines", *Proc. 27th Int. Conf. Mach. Learn.*, Haifa, Israel, 2010.

[173] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling", 2014. arXiv: 1412.3555 [cs.NE].

[174] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need", *31st Conf. Neural Inf. Process. Syst. (NIPS 2017)*, Long Beach, CA, USA, 2017. arXiv: 1706.03762 [cs.CL].

[175] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 2019. arXiv: 1810.04805v2 [cs.CL].

[176] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate", *3rd Int. Conf. Learn. Represent. (ICLR 2015)*, San Diego, CA, USA, 2015. arXiv: 1409.0473 [cs.CL].

[177] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, 2020. DOI: 10.1109/TPAMI.2018.2858826.

[178] C. Peyron, J. Faraco, W. Rogers, B. Ripley, S. Overeem, Y. Charnay, S. Nevsimalova, M. Aldrich, D. Reynolds, R. Albin, R. Li, M. Hungs, M. Pedrazzoli, M. Padigaru, M. Kucherlapati, J. Fan, R. Maki, G. J. Lammers, C. Bouras, R. Kucherlapati, S. Nishino, and E. Mignot, "A mutation in a case of early onset narcolepsy and a generalized absence of hypocretin peptides in human narcoleptic brains", *Nat. Med.*, vol. 6, no. 9, pp. 991–997, 2000. DOI: 10.1038/79690.

[179] E. Mignot, G. J. Lammers, B. Ripley, M. Okun, S. Nevsimalova, S. Overeem, J. Vankova, J. Black, J. Harsh, C. Bassetti, H. Schrader, and S. Nishino, "The Role of Cerebrospinal Fluid Hypocretin Measurement in the Diagnosis of Narcolepsy and Other Hypersomnias", *Arch. Neurol.*, vol. 59, no. 10, p. 1553, 2002. DOI: 10.1001/archneur.59.10.1553.

[180] B. R. Kornum, "Narcolepsy type 1: what have we learned from immunology?", *Sleep*, pp. 1–5, 2020. DOI: 10.1093/sleep/zsaa055.

[181] R. Fronczek, I. Arnulf, C. R. Baumann, K. Maski, F. Pizza, and L. M. Trotti, "To split or to lump? Classifying the central disorders of hypersomnolence", *Sleep*, zsaa044, 2020. DOI: 10.1093/sleep/zsaa044.

[182] F. Han, L. Lin, B. Schormair, F. Pizza, G. Plazzi, H. M. Ollila, S. Nevsimalova, P. Jennum, S. Knudsen, J. Winkelmann, C. Coquillard, F. Babrzadeh, T. M. Strom, C. Wang, M. Mindrinos, M. F. Vina, and E. Mignot, "HLA DQB1*06:02 Negative Narcolepsy with Hypocretin/Orexin Deficiency", *Sleep*, vol. 37, no. 10, pp. 1601–1608, 2014. DOI: 10.5665/sleep.4066.

[183] M. R. Littner, C. Kushida, M. Wise, D. G. Davila, T. Morgenthaler, T. Lee-Chiong, M. Hirshkowitz, D. L. Loube, D. Bailey, R. B. Berry, S. Kapen, and M. Kramer, "Practice Parameters for Clinical Use of the Multiple Sleep Latency Test and the Maintenance of Wakefulness Test", *Sleep*, vol. 28, no. 1, pp. 113–121, 2005. DOI: 10.1093/sleep/28.1.113.

[184] O. Andlauer, H. Moore, S.-C. Hong, Y. Dauvilliers, T. Kanbayashi, S. Nishino, F. Han, M. H. Silber, T. Rico, M. Einen, B. R. Kornum, P. Jennum, S. Knudsen, S. Nevsimalova, F. Poli, G. Plazzi, and E. Mignot, "Predictors of Hypocretin (Orexin) Deficiency in Narcolepsy Without Cataplexy", *Sleep*, vol. 35, no. 9, pp. 1247–1255, 2012. DOI: 10.5665/sleep.2080.

[185] G. Luca, J. Haba-Rubio, Y. Dauvilliers, G.-J. Lammers, S. Overeem, C. E. Donjacour, G. Mayer, S. Javidi, A. Iranzo, J. Santamaria, R. Peraita-Adrados, H. Hor, Z. Kutalik, G. Plazzi, F. Poli, F. Pizza, I. Arnulf, M. Lecendreux, C. Bassetti, J. Mathis, R. Heinzer, P. Jennum, S. Knudsen, P. Geisler, A. Wierzbicka, E. Feketeova, C. Pfister, R. Khatami, C. Baumann, and M. Tafti, "Clinical, polysomnographic and genome-wide association analyses of narcolepsy with cataplexy: a European Narcolepsy Network study", *J. Sleep Res.*, vol. 22, no. 5, pp. 482–495, 2013. DOI: 10.1111/jsr.12044.

[186] Y. Dauvilliers, A. Gosselin, J. Paquet, J. Touchon, M. Billiard, and J. Montplaisir, "Effect of age on MSLT results in patients with narcolepsy-cataplexy", *Neurology*, vol. 62, no. 1, pp. 46–50, 2004. DOI: 10.1212/01.WNL.0000101725.34089.1E.

[187] A. Moscovitch, M. Partinen, and C. Guilleminault, “The positive diagnosis of narcolepsy and narcolepsy’s borderland”, *Neurology*, vol. 43, no. 1 Part 1, pp. 55–55, 1993. DOI: 10.1212/WNL.43.1_Part_1.55.

[188] B. Frauscher, L. Ehrmann, T. Mitterling, D. Gabelia, V. Gschliesser, E. Brandauer, W. Poewe, and B. Högl, “Delayed Diagnosis, Range of Severity, and Multiple Sleep Comorbidities: A Clinical and Polysomnographic Analysis of 100 Patients of the Innsbruck Narcolepsy Cohort”, *J. Clin. Sleep Med.*, vol. 09, no. 08, pp. 805–812, 2013. DOI: 10.5664/jcsm.2926.

[189] International Xyrem Study Group, “A double-blind, placebo-controlled study demonstrates sodium oxybate is effective for the treatment of excessive daytime sleepiness in narcolepsy”, *J. Clin. Sleep Med.*, vol. 1, no. 4, pp. 391–397, 2005.

[190] F. Pizza, S. Vandi, M. Iloti, C. Franceschini, R. Liguori, E. Mignot, and G. Plazzi, “Nocturnal Sleep Dynamics Identify Narcolepsy Type 1”, *Sleep*, vol. 38, no. 8, pp. 1277–1284, 2015. DOI: 10.5665/sleep.4908.

[191] J. A. Christensen, L. Kempfner, H. L. Leonthin, M. Hvidtfelt, M. Nikolic, B. R. Kornum, and P. Jennum, “Novel method for evaluation of eye movements in patients with narcolepsy”, *Sleep Med.*, vol. 33, pp. 171–180, 2017. DOI: 10.1016/j.sleep.2016.10.016.

[192] J. A. E. Christensen, O. Carrillo, E. B. Leary, P. E. Peppard, T. Young, H. B. D. Sorensen, P. Jennum, and E. Mignot, “Sleep-stage transitions during polysomnographic recordings as diagnostic features of type 1 narcolepsy”, *Sleep Med.*, vol. 16, no. 12, pp. 1558–1566, 2015. DOI: 10.1016/j.sleep.2015.06.007.

[193] T. Roth, Y. Dauvilliers, E. Mignot, J. Montplaisir, J. Paul, T. Swick, and P. Zee, “Disrupted Nighttime Sleep in Narcolepsy”, *J. Clin. Sleep Med.*, vol. 09, no. 09, pp. 955–965, 2013. DOI: 10.5664/jcsm.3004.

[194] M. H. Hansen, B. R. Kornum, and P. Jennum, “Sleep–wake stability in narcolepsy patients with normal, low and unmeasurable hypocretin levels”, *Sleep Med.*, vol. 34, pp. 1–6, 2017. DOI: 10.1016/j.sleep.2017.01.021.

[195] P. Drakatos, C. A. Kosky, S. E. Higgins, R. T. Muza, A. J. Williams, and G. D. Leschziner, “First rapid eye movement sleep periods and sleep-onset rapid eye movement periods in sleep-stage sequencing of hypersomnias”, *Sleep Med.*, vol. 14, no. 9, pp. 897–901, 2013. DOI: 10.1016/j.sleep.2013.03.021.

[196] Y. Liu, J. Zhang, V. Lam, C. K. W. Ho, J. Zhou, S. X. Li, S. P. Lam, M. W. M. Yu, X. Tang, and Y.-K. Wing, “Altered Sleep Stage Transitions of REM Sleep: A Novel and Stable Biomarker of Narcolepsy”, *J. Clin. Sleep Med.*, vol. 11, no. 08, pp. 885–894, 2015. DOI: 10.5664/jcsm.4940.

[197] I. Guyon, J. Weston, and S. Barnhill, “Gene Selection for Cancer Classification using Support Vector Machines”, *Mach. Learn.*, vol. 46, pp. 389–422, 2002. DOI: 10.1023/A:1012487302797.

[198] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: The MIT Press, 2006, ISBN: ISBN 0-262-18253-X.

[199] J. Hensman, A. Matthews, and Z. Ghahramani, “Scalable Variational Gaussian Process Classification”, *Proc. 18th Int. Conf. Artif. Intell. Stat.*, San Diego, CA, USA: JMLR: W&CP, 2015. arXiv: 1411.2005 [stat.ML].

[200] A. G. D. G. Matthews, T. Nickson, A. Boukouvalas, and J. Hensman, “GPflow: A Gaussian Process Library using TensorFlow”, *J. Mach. Learn. Res.*, vol. 18, pp. 1–6, 2017. arXiv: 1610.08733 [stat.ML].

[201] A. V. Olsen, J. Stephansen, E. Leary, P. E. Peppard, H. Sheungshul, P. J. Jennum, H. Sorensen, and E. Mignot, “Diagnostic value of sleep stage dissociation as visualized on a 2-dimensional sleep state space in human narcolepsy”, *J. Neurosci. Methods*, vol. 282, pp. 9–19, 2017. DOI: 10.1016/j.jneumeth.2017.02.004.

[202] J. B. Jensen, H. B. D. Sorensen, J. Kempfner, G. L. Sørensen, S. Knudsen, and P. Jennum, “Sleep–Wake Transition in Narcolepsy and Healthy Controls Using a Support Vector Machine”, *J. Clin. Neurophysiol.*, vol. 31, no. 5, pp. 397–401, 2014. DOI: 10.1097/WNP.0000000000000074.

[203] A. Vassalli, J. M. Dellepiane, Y. Emmenegger, S. Jimenez, S. Vandi, G. Plazzi, P. Franken, and M. Tafti, “Electroencephalogram paroxysmal theta characterizes cataplexy in mice and children”, *Brain*, vol. 136, no. 5, pp. 1592–1608, 2013. DOI: 10.1093/brain/awt069.

[204] J. Reiter, E. Katz, T. E. Scammell, and K. Maski, “Usefulness of a Nocturnal SOREMP for Diagnosing Narcolepsy with Cataplexy in a Pediatric Population”, *Sleep*, vol. 38, no. 6, pp. 859–65, 2015. DOI: 10.5665/sleep.4728.

[205] A. B. A. Stevner, D. Vidaurre, J. Cabral, K. Rapuano, S. F. V. Nielsen, E. Tagliazucchi, H. Laufs, P. Vuust, G. Deco, M. W. Woolrich, E. Van Someren, and M. L. Kringelbach, “Discovery of key whole-brain transitions and dynamics during human wakefulness and non-REM sleep”, *Nat. Commun.*, vol. 10, no. 1, p. 1035, 2019. DOI: 10.1038/s41467-019-08934-3.

[206] J. A. E. Christensen, R. Wassing, Y. Wei, J. R. Ramautar, O. Lakbila-Kamal, P. J. Jennum, and E. J. Van Someren, “Data-driven analysis of EEG reveals concomitant superficial sleep during deep sleep in insomnia disorder”, *Front. Neurosci.*, vol. 13, p. 598, 2019. DOI: 10.3389/fnins.2019.00598.

[207] H. Koch, P. Jennum, and J. A. E. Christensen, “Automatic sleep classification using adaptive segmentation reveals an increased number of rapid eye movement sleep transitions”, *J. Sleep Res.*, vol. 28, e12780, 2019. DOI: 10.1111/jsr.12780.

[208] G. L. Sorensen, S. Knudsen, and P. Jennum, “Sleep Transitions in Hypocretin-Deficient Narcolepsy”, *Sleep*, vol. 36, no. 8, pp. 1173–1177, 2013. DOI: 10.5665/sleep.2880.

[209] J. A. E. Christensen, P. Jennum, H. Koch, R. Frandsen, M. Zoetmulder, L. Arvastson, S. R. Christensen, and H. B. D. Sorensen, “Sleep stability and transitions in patients with idiopathic REM sleep behavior disorder and patients with Parkinson’s disease”, *Clin. Neurophysiol.*, vol. 127, no. 1, pp. 537–543, 2016. DOI: 10.1016/j.clinph.2015.03.006.

[210] Y. Dauvilliers, L. Barateau, R. Lopez, A. L. Rassu, S. Chenini, S. Beziat, and I. Jaussent, "Narcolepsy Severity Scale: a reliable tool assessing symptom severity and consequences", *Sleep*, pp. 1–11, 2020. DOI: 10.1093/sleep/zsaa009.

[211] A. R. Schwartz, M. Cohen-Zion, L. V. Pham, A. Gal, M. Sowho, F. P. Sgambati, T. Klopfer, M. A. Guzman, E. M. Hawks, T. Etzioni, L. Glasner, E. Druckman, and G. Pillar, "Brief Digital Sleep Questionnaire Powered by Machine Learning Prediction Models Identifies Common Sleep Disorders", *Sleep Med.*, 2020. DOI: 10.1016/j.sleep.2020.03.005, *in press*.

[212] G. J. Lammers, C. L. Bassetti, L. Dolenc-Groselj, P. J. Jennum, U. Kallweit, R. Khatami, M. Lecendreux, M. Manconi, G. Mayer, M. Partinen, G. Plazzi, P. J. Reading, J. Santamaria, K. Sonka, and Y. Dauvilliers, "Diagnosis of central disorders of hypersomnolence: A reappraisal by European experts", *Sleep Med. Rev.*, vol. 52, p. 101 306, 2020. DOI: 10.1016/j.smrv.2020.101306.

[213] 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, D. B. Kirsch, R. K. Malhotra, J. L. Martin, E. J. Olson, K. Ramar, C. L. Rosen, J. A. Rowley, and A. V. Shelgikar, "Artificial intelligence in sleep medicine: an American Academy of Sleep Medicine position statement", *J. Clin. Sleep Med.*, vol. 16, no. 4, pp. 605–607, 2020. DOI: 10.5664/jcsm.8288.

[214] C. A. Goldstein, R. B. Berry, D. T. Kent, D. A. Kristo, A. A. Seixas, S. Redline, and M. B. Westover, "Artificial intelligence in sleep medicine: background and implications for clinicians", *J. Clin. Sleep Med.*, vol. 16, no. 4, pp. 609–618, 2020. DOI: 10.5664/jcsm.8388.

[215] D. C. Lim, D. R. Mazzotti, K. Sutherland, J. W. Mindel, J. Kim, P. A. Cistulli, U. J. Magalang, A. I. Pack, P. de Chazal, and T. Penzel, "Reinventing polysomnography in the age of precision medicine", *Sleep Med. Rev.*, vol. 52, p. 101 313, 2020. DOI: 10.1016/j.smrv.2020.101313.

[216] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding", 2018. arXiv: 1807.03748 [cs.LG].

[217] O. J. Hénaff, A. Srinivas, J. De Fauw, A. Razavi, C. Doersch, S. M. A. Eslami, and A. van den Oord, "Data-Efficient Image Recognition with Contrastive Predictive Coding", 2019. arXiv: 1905.09272 [cs.CV].

[218] H. Banville, I. Albuquerque, A. Hyvarinen, G. Moffat, D.-A. Engemann, and A. Gramfort, "Self-Supervised Representation Learning from Electroencephalography Signals", *2019 IEEE 29th Int. Work. Mach. Learn. Signal Process.*, Pittsburgh, PA, USA: IEEE, 2019, pp. 1–6. DOI: 10.1109/MLSP.2019.8918693.

[219] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds., *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, ser. Lecture Notes in Computer Science. Cham: Springer, Cham, 2019, vol. 11700, p. 435. DOI: 10.1007/978-3-030-28954-6.

Part IV

APPENDIX

A

PAPER I

TITLE: Deep residual networks for automatic sleep stage classification of raw polysomnographic waveforms

AUTHORS: Alexander Neergaard Olesen, Poul Jennum, Paul Peppard, Emmanuel Mignot, Helge Bjarup Dissing Sorensen

CONFERENCE: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)

STATUS: Published

DOI: 10.1109/EMBC.2018.8513080

FULL CITATION: A. N. Olesen, P. Jennum, P. Peppard, E. Mignot and H. B. D. Sorensen, "Deep residual networks for automatic sleep stage classification of raw polysomnographic waveforms," *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Honolulu, HI, 2018, pp. 1-4.

Deep residual networks for automatic sleep stage classification of raw polysomnographic waveforms

Alexander N. Olesen[†,1,2], Poul Jennum[3], Paul Peppard[4], Emmanuel Mignot[2], and Helge B. D. Sorensen[1],

***Abstract*— We have developed an automatic sleep stage classification algorithm based on deep residual neural networks and raw polysomnogram signals. Briefly, the raw data is passed through 50 convolutional layers before subsequent classification into one of five sleep stages. Three model configurations were trained on 1850 polysomnogram recordings and subsequently tested on 230 independent recordings. Our best performing model yielded an accuracy of 84.1% and a Cohen's kappa of 0.746, improving on previous reported results by other groups also using only raw polysomnogram data. Most errors were made on non-REM stage 1 and 3 decisions, errors likely resulting from the definition of these stages. Further testing on independent cohorts is needed to verify performance for clinical use.**

I. Introduction

Sleep staging is the principal tool available to medical doctors in the analysis of sleep disorders. Natural human sleep consists of recurring cycles of three to four distinct phases, which are primarily characterized by changes in brain activity, eye movements, muscle activations and breathing. A polysomnogram (PSG) containing electroencephalography (EEG), electrooculography (EOG), electromyography (EMG) and other signals is collected during sleep, and subsequently processed and analyzed by sleep technicians according to standards by the American Academy of Sleep Medicine (AASM). Each 30 s epoch of data is categorized into either wakefulness (W), rapid eye movement (REM) sleep, or one of three stages of non-REM sleep (N1, N2, N3) [1]. However, this approach is prone to subjective interpretation of sleep staging rules, which have prompted extensive research in using various signal processing and machine learning approaches [2].

Attempts at exploiting deep learning models for sleep staging have been proposed recently. One group used a transfer learning-approach to characterize sleep stages [3], where 30 s epochs of Fpz-Cz EEG were subjected to multitaper spectral estimation (MTSE) in order to create spectral image representations [4], that ultimately were fed as input to a VGG-16 model stripped of the last layers [5]. This approach was cross-validated using a leave-one-out scheme on 20 subjects and yielded a bootstrapped accuracy of $86\% \pm 2\%$.

Using MTSE for representing EEG data was also investigated in [6], where the authors compared various machine and deep learning models trained on either raw EEG waveforms, MTSE spectrograms, or 96 expert-defined features. They tested their best performing model on recordings from 1000 individual subjects and obtained an accuracy of $85.76\,\%$ and a Cohen's kappa of 0.79 using a combination of expert-defined features and recurrent neural networks (RNN). On the same test set, they obtained accuracy/kappa values of $77.31\,\%$ and 0.71 using a deep learning model trained on raw EEG waveforms.

However, it is still unclear whether manual feature extraction such as sleep spindle/K-complex detection, or data transformations, such as spectrograms or MTSE, are strictly necessary for efficient deep learning, and there is still room for improvement in the current state of the art for raw PSG analysis.

We propose a novel method for automatic sleep staging combining state of the art deep learning networks with raw PSG data to accurately capture the complex relationships found in PSG data without resorting to data transformations and manual feature engineering.

II. Data

A database containing 2310 recordings extracted from the Wisconsin Sleep Cohort was used in this study. Specific acquisition details concerning the PSGs are described in [7]. The entire set of PSG studies was randomly split into training (train), validation (eval), and testing (test) subgroups in an 8:1:1 ratio. Detailed demographic information as well as relevant PSG variables for all three subgroups are provided in table I including apnea-hypopnea index (AHI) and time spent in each sleep stage based on manual scoring.

III. Methods

A. Signal extraction and pre-processing

Central and occipital EEG from right hemisphere, left and right EOG, and chin EMG channels were extracted from each PSG study. To accommodate different equipment setups used for recording studies, each channel was upsampled to 200 Hz. Following resampling, signals were filtered using zero-phase Butterworth filters with frequency ranges recommended by the AASM [1]. Since dynamic ranges vary considerably across channels, each signal was soft-normalized using the 5th and 95th quantiles, such that

$$\mathbf{x}_{\text{norm}} = 2\,\frac{\mathbf{x} - \mathrm{Q}_{0.05}(\mathbf{x})}{\mathrm{Q}_{0.95}(\mathbf{x}) - \mathrm{Q}_{0.05}(\mathbf{x})} - 1, \tag{1}$$

[†] Corresponding author: aneol@elektro.dtu.dk.

[1] Department of Electrical Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark.

[2] Stanford Center for Sleep Sciences and Medicine, Stanford University, Palo Alto, CA, USA.

[3] Danish Center for Sleep Medicine, Department of Neurophysiology, Rigshospitalet, Glostrup, Denmark.

[4] University of Wisconsin School of Medicine and Public Health, Madison, WI, USA.

978-1-5386-3646-6/18/$31.00 ©2018 IEEE

TABLE I

EXTRACTED WSC COHORT DEMOGRAPHICS FOR EACH SUBGROUP. SIGNIFICANT p-VALUES ARE HIGHLIGHTED IN BOLD.

	Train	Eval	Test	p-value
n (male)	1850 (1010)	230 (112)	230 (120)	0.210
Age (years)	59.2 ± 8.4	59.9 ± 8.5	60.4 ± 8.2	0.092
BMI ($\mathrm{kg\,m^{-2}}$)	31.7 ± 7.2	31.0 ± 6.9	32.2 ± 7.7	0.203
AHI ($\mathrm{h^{-1}}$)	12.6 ± 15.6	11.5 ± 14.9	12.4 ± 16.2	0.600
PSG dur. (h)	7.4 ± 0.8	7.4 ± 0.7	7.4 ± 0.8	0.947
W (%)	18.5 ± 11.3	17.2 ± 11.1	19.6 ± 11.8	0.071
N1 (%)	8.2 ± 4.5	8.8 ± 5.6	8.9 ± 5.1	**0.038**
N2 (%)	54.2 ± 10.3	54.0 ± 10.9	52.4 ± 11.0	**0.048**
N3 (%)	5.8 ± 6.4	6.4 ± 7.0	6.0 ± 7.0	0.433
REM (%)	13.3 ± 5.9	13.7 ± 5.8	13.2 ± 5.7	0.635

where $\mathbf{x}_{\mathrm{norm}}$ denotes the normalized version of the signal $\mathbf{x}$, and $\mathrm{Q}_{0.05}(\mathbf{x})$ and $\mathrm{Q}_{0.95}(\mathbf{x})$ denotes the 5th and 95th percentile, respectively. Doubling and subtracting by one rescales $\mathrm{Q}_{0.05}(\mathbf{x})$ and $\mathrm{Q}_{0.95}(\mathbf{x})$ to -1 and 1, respectively.

Finally, each signal was segmented into 30 s epochs corresponding to AASM criteria [1], resulting in a tensor $\mathbf{X}$ with elements

$$(x_{nc\cdot t}) \in \mathbb{R}^{N \times C \times 1 \times T}, \tag{2}$$

with $N = 16$, $C = 5$, and $T = 6000$ being batch size, number of signals, and number of timesteps for one epoch, respectively.[1]

B. Deep residual network model

We applied a deep learning model inspired by the residual network models proposed in [8], [9]. These types of models employ residual skip connections between layers in order to maintain a proper gradient backpropagation through the network. This feature allows for extremely deep network structures, and a specific variant of this model with 152 layers came in 1st place in the ILSVRC '15 image classification competition [8].

1) Architecture: The residual network model is illustrated in Fig. 1. Briefly, the bulk network comprised 50 convoluational (conv) and dense layers arranged in four block layers of four bottlenecked residual blocks each.

A single bottleneck residual block contains three triplets of a batch normalization layer, a rectified linear unit (ReLU) activation layer, and a conv layer. This pre-activation configuration has shown benefits with regards to trainability and generalization compared to vanilla residual blocks [9]. Projection shortcuts were used between the first ReLU and conv layers to the output of the last conv layer. Kernel sizes were set to 1×1 for the first and third conv layers, and 1×3 for the second conv layer. The number of output filters for each residual block was $l \times f$ with l being the block layer index and $f = 16$, resulting in a total of 256 filters after the final conv layer.

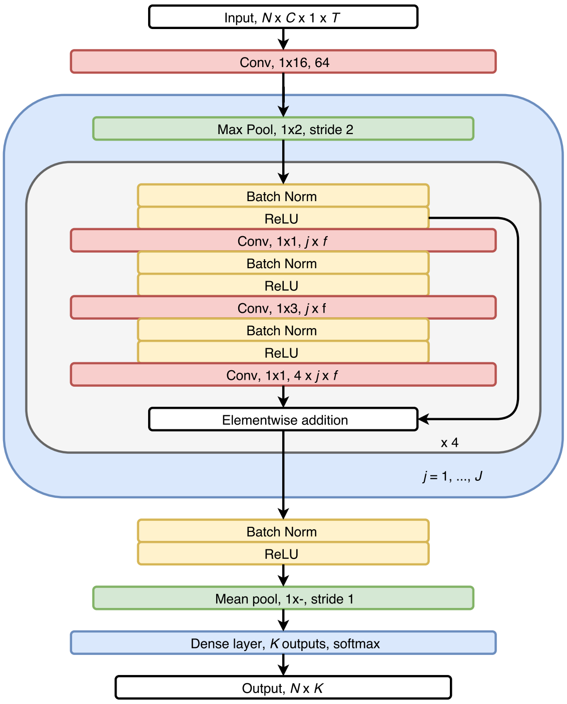


Fig. 1. Model architecture. The input tensor has shape $(N, C, 1, T)$, where N, C, T correspond to the batch size, number of signals, and length of each 30 s epoch, respectively. The output tensor has shape $N \times K$ with $K = 5$ sleep stages, while $J = 4$, and $f = 16$ is the number of block layers and base number of filters.

Before the bottleneck blocks, the input tensor $\mathbf{X}$ was passed through an initial conv layer consisting of 64 1×16 filters, and then through a maximum pooling (max pool) layer with a 1×2 kernel and stride size, effectively reducing the time-resolution by a factor of 2. This max pool operation was implemented in the beginning of each block layer.

The output tensor from the block layers was subsequently passed to a final batch normalization and ReLU activation layer, followed by a mean pooling layer to reduce the tensor to $\mathbf{X} = (x_{nk}) \in \mathbb{R}^{N \times 256}$. Finally, a fully connected layer with $K = 5$ output units corresponding to the sleep stages resulted in the following output tensor

$$\mathbf{P} = (p_{nk}) \in \mathbb{R}^{N \times K}, \quad p_{nk} = \frac{\exp z_{nk}}{\sum_k^K \exp z_{nk}} \tag{3}$$

with p_{nk} containing the softmax activations of the output units z_{nk} from the fully connected layer for the nth subject and the kth sleep stage. The predicted class for the nth subject can then be calculated as

$$\hat{y}_n = \arg\max_k p_{nk}. \tag{4}$$

2) Training: The optimization problem was constructed using cross entropy loss across K classes and N epochs as objective function, such that

$$\mathcal{L}(\mathbf{p}_n | \mathbf{y}_n, \mathrm{W}) = -\sum_{k=1}^{K} y_{nk} \log p_{nk}, \tag{5}$$

[1] The 1-dimensional convolution `tf.layers.conv1d` reshapes the input argument to subsequently call `tf.layers.conv2d` in TensorFlow. To reduce computational costs, we introduce a singleton dimension.

is the calculated cross entropy loss for epoch n given predicted class probabilities $\mathbf{p}_n$, true class labels $\mathbf{y}_n$, and the set of current weights W. Then, the average cost across a batch of data is

$$\mathcal{C}(\mathbf{P}|\,\mathbf{Y},\mathsf{W}) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}(\mathbf{p}_n|\mathbf{y}_n,\mathsf{W}). \tag{6}$$

The cost function was optimized using the Adam optimization algorithm with default hyperparameters [10]. Weights were initialized using variance scaling [11], and we applied weight decay during training with $\lambda = 10^{-4}$. The initial learning rate was set to $\alpha = 10^{-3}$ and was multiplied by 0.1 every 50000 steps.

In order to investigate the effect of the imbalanced data on the network performance, we trained the following three different configurations. First, we defined a *baseline* configuration as described in the previous sections. The second was a *weighted* configuration, where the cost function in eq. (6) was replaced with an average weighted by the inverse frequency for the correct class, such that

$$\mathcal{C}(\hat{\mathbf{Y}}|\,\mathbf{Y},\mathsf{W}) = \frac{\sum_n^N \omega_n(\mathbf{y}_n)\mathcal{L}(\hat{\mathbf{y}}_n|\mathbf{y}_n,\mathsf{W})}{\sum_n^N \omega_n(\mathbf{y}_n)}, \tag{7}$$

where $\omega_n(\mathbf{y}_n)$ is the inverse frequency for the correct class for the nth subject in the current batch. Finally, a *weighted* configuration was tested, in which we performed resampling of the training dataset in order to balance classes. We oversampled the N1, N3, and REM classes with replacement, while undersampling the N2 class in order to have approximately equal fractions of each class in total.

Models were implemented in TensorFlow 1.4, and trained on a single workstation running Ubuntu 16.04 with a Ryzen 7 1700X 8-core CPU, an NVIDIA GTX 1080 Ti GPU with 11 GB memory, and 32 GB RAM memory.

C. Performance metrics

Individual precision, recall and F1 scores (Pr, Re, F1) were calculated for each sleep stage and subsequently aggregated for each recording by stage frequency weighting, such that

$$\mathrm{Pr}_{nk} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad \mathrm{Pr}_n = \frac{\sum_k \beta_{nk}\mathrm{Pr}_{nk}}{\sum_k \beta_{nk}} \tag{8}$$

$$\mathrm{Re}_{nk} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad \mathrm{Re}_n = \frac{\sum_k \beta_{nk}\mathrm{Re}_{nk}}{\sum_k \beta_{nk}} \tag{9}$$

$$\mathrm{F1}_{nk} = 2\cdot\frac{\mathrm{Pr}_{nk}\cdot\mathrm{Re}_{nk}}{\mathrm{Pr}_{nk}+\mathrm{Re}_{nk}}, \quad \mathrm{F1}_n = \frac{\sum_k \beta_{nk}\mathrm{F1}_{nk}}{\sum_k \beta_{nk}}, \tag{10}$$

where β_{nk} is the frequency of stage k for recording n, and TP, FP and FN are true positives, false positives, and false negatives, respectively. Overall accuracy (Acc) and Cohen's kappa (κ) were also calculated for each recording. All metrics were summarized by mean and standard deviations.

D. Statistical tests

Demographic and PSG variables were tested with ANOVAS after establishing normality, while gender was tested with a χ^2 test. Significance was set at $p = 0.05$.

TABLE II
AVERAGED PERFORMANCE METRICS FOR CONFIGURATIONS ACROSS TRAIN AND EVAL SUBGROUPS WITH BEST SHOWN IN BOLD.

		baseline	weighted	balanced
Train	Acc (%)	$\mathbf{86.1 \pm 5.5}$	79.4 ± 7.1	80.4 ± 7.3
	κ (%)	$\mathbf{77.1 \pm 8.6}$	69.5 ± 9.7	70.7 ± 9.8
	Pr (%)	87.1 ± 4.9	88.7 ± 4.1	$\mathbf{88.9 \pm 4.0}$
	Re (%)	$\mathbf{86.1 \pm 5.5}$	79.4 ± 7.1	80.4 ± 7.3
	F1 (%)	$\mathbf{85.3 \pm 6.1}$	81.8 ± 6.6	82.6 ± 6.9
Eval	Acc (%)	$\mathbf{85.0 \pm 6.1}$	78.4 ± 7.3	79.7 ± 7.4
	κ (%)	$\mathbf{75.4 \pm 9.5}$	68.1 ± 10.5	69.7 ± 10.0
	Pr (%)	86.3 ± 5.3	87.8 ± 4.8	$\mathbf{88.0 \pm 4.9}$
	Re (%)	$\mathbf{85.0 \pm 6.1}$	78.4 ± 7.3	79.7 ± 7.4
	F1 (%)	$\mathbf{84.0 \pm 7.2}$	80.7 ± 7.1	81.9 ± 7.1

TABLE III
AGGREGATED CONFUSION MATRIX AND STAGE-SPECIFIC PERFORMANCE METRICS IN TEST SUBGROUP.

	W	N1	N2	N3	REM	Pr (%)	Re (%)	F1 (%)
W	37980	1322	852	2	327	84.3	93.8	88.8
N1	3922	8784	3545	0	2193	51.9	47.6	49.7
N2	1756	5136	99564	1091	991	88.6	91.7	90.2
N3	18	1	7932	4063	14	78.8	33.8	47.3
R	1361	1680	465	0	23931	87.2	87.2	87.2

IV. RESULTS AND DISCUSSION

Performance metrics for the train and eval subgroups are shown in table II. Not accounting for Pr, the baseline configuration compares favorably to the weighted and balanced configurations on both subgroups with an average accuracy of 85.0 % and a Cohen's kappa of 75.4 on the eval subgroup. Since the training data is imbalanced in favor of N2, it would be fair to assume overfitting to the majority class, however, the lower spread in both precision and recall does not support this. Evaluating the baseline model on the test subgroup gave only a slight drop in accuracy and κ, indicating that the model generalizes well, see table III and table IV. The lowest sensitivity is obtained for N1 and N3, which is in accordance with clinical experience reported in the literature [12]–[15]. N1 is a transitional stage between wakefulness, drowsiness and sleep often containing beta and alpha activity in epochs of low interscorer agreement, which explains the low predictive power in the confusion matrix. The sleep continuum is also apparent in Fig. 2 which shows the manually and automatically scored hypnograms in the middle and bottom traces, and the hypnodensity graph in the top trace for a representative subject in the test subgroup. The hypnodensity is a probabilistic representation of the hypnogram, which has found use in the detection of Parkinson's and narcolepsy [16]–[18]. Our baseline model attains favorable performance when comparing to the results reported for the raw waveform CNN model in [6] with both higher accuracy and Cohen's kappa. However, it should be

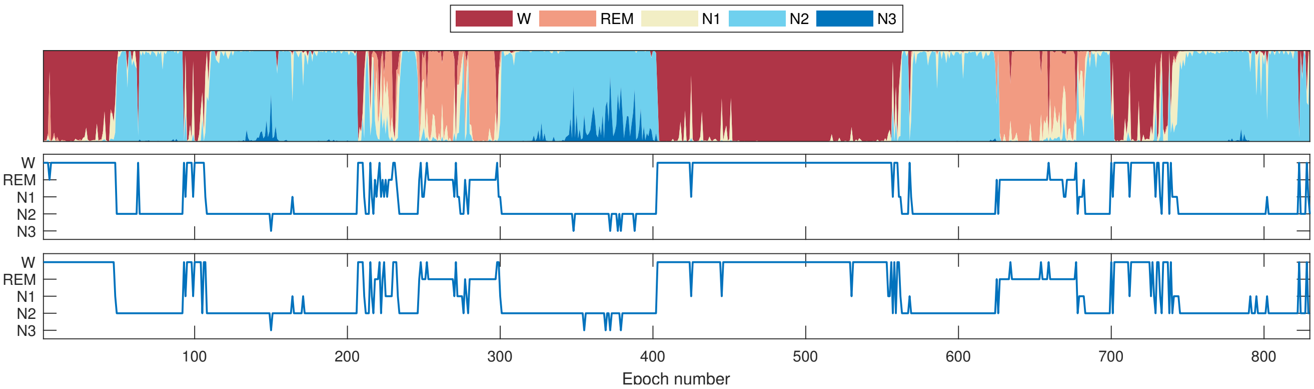


Fig. 2. Top: hypnodensity graph of per-epoch probability distributions, middle: automatically scored hypnogram by applying eq. (4), bottom: manually scored hypnogram. Note the intrusions of N3 into N2 around epoch 150 and 370, and N1 into W around 420.

TABLE IV

PERFORMANCE ACROSS RECORDINGS IN TEST SUBGROUP.

Acc	κ	Pr	Re	F1
84.1 ± 6.9	0.746 ± 0.099	85.7 ± 6.1	84.1 ± 6.9	83.1 ± 7.6

stressed that [6] used EEG from 9000 recordings, while our model uses EEG, EOG and EMG from 1850 recordings. Furthermore, our baseline model performs only slightly worse compared to the best-performing model using manual feature engineering and RNNs in [6]. This indicates a possible performance gain by adding recurrent networks, such as long short-term memory cells, to our network.

A possible limiting factor to our model is the filter kernels. The small filter sizes in block layers might not be able to accurately capture the physiological dynamics, but there are indications that many, smaller kernels are preferable to fewer, larger kernels when comparing model complexity versus computational costs [19].

Future work will include adding more data to balance classes, and adding long short-term memory cells to the network in order to model temporal dynamics between epochs. As we performed minimal hyperparameter tuning in this work, investigating the effects of changing the network specifications to optimize performance is also a relevant area of future research.

V. Conclusion

We have shown that common data transformations such as spectrograms are not necessary for automatic sleep staging. Combining residual learning networks and raw PSG waveforms, we obtained an average accuracy of 84.1% and Cohen's kappa of 0.745, improving on previously reported results on raw PSG sleep staging. Further testing on independent cohorts will illuminate the clinical applicability of this method, while introducing more data and memory cells will be explored to increase performance even further.

References

[1] Berry *et al.*, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.3*. Darien, Illinois: The American Academy of Sleep Medicine, 2016.

[2] Şen *et al.*, "A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms," *J. Med. Syst.*, vol. 38, no. 3, 2014.

[3] Vilamala, Madsen, and Hansen, "Deep Convolutional Neural Networks for Interpretable Analysis of EEG Sleep Stage Scoring," in *IEEE Int. Work. Mach. Learn. Signal Process. Sept. 25-28*, Tokyo, Japan, 2017.

[4] Thomson, "Spectrum estimation and harmonic analysis," *Proc. IEEE*, vol. 70, no. 9, pp. 1055–1096, 1982.

[5] Simonyan and Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[6] Biswal *et al.*, "SLEEPNET: Automated Sleep Staging System via Deep Learning," pp. 1–17, 2017. [Online]. Available: http://arxiv.org/abs/1707.08262

[7] Young *et al.*, "Sleep Disordered Breathing and Mortality: Eighteen-Year Follow-up of the Wisconsin Sleep Cohort," *Sleep*, vol. 31, no. 8, pp. 291–292, 2008.

[8] He *et al.*, "Deep Residual Learning for Image Recognition," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[9] He *et al.*, "Identity Mappings in Deep Residual Networks," in *Comput. Vis. – ECCV 2016*, vol. abs/1603.0, 2016, pp. 630–645.

[10] Kingma and Ba, "Adam: A Method for Stochastic Optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2014, pp. 1–15.

[11] He *et al.*, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *2015 IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[12] Younes, "The case for using digital EEG analysis in clinical sleep medicine," *Sleep Sci. Pract.*, vol. 1, no. 1, p. 2, 2017.

[13] Rosenberg and Van Hout, "The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring," *J. Clin. Sleep Med.*, vol. 9, no. 1, pp. 81–87, 2013.

[14] Norman *et al.*, "Interobserver agreement among sleep scorers from different centers in a large dataset." *Sleep*, vol. 23, no. 7, pp. 901–8, 2000.

[15] Younes, Raneri, and Hanly, "Staging sleep in polysomnograms: Analysis of inter-scorer variability," *J. Clin. Sleep Med.*, vol. 12, no. 6, pp. 885–894, 2016.

[16] Koch *et al.*, "Automatic sleep classification using a data-driven topic model reveals latent sleep states." *J. Neurosci. Methods*, vol. 235, pp. 130–137, 2014.

[17] Christensen *et al.*, "Data-driven modeling of sleep EEG and EOG reveals characteristics indicative of pre-Parkinson's and Parkinson's disease," *J. Neurosci. Methods*, vol. 235, pp. 262–276, 2014.

[18] Stephansen *et al.*, "The use of neural networks in the analysis of sleep stages and the diagnosis of narcolepsy," oct 2017. [Online]. Available: http://arxiv.org/abs/1710.02094

[19] Szegedy *et al.*, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, jun 2016, pp. 2818–2826.

B

PAPER II

TITLE: Automatic sleep stage classification with deep residual networks in a mixed-cohort setting

AUTHORS: Alexander Neergaard Olesen, Poul Jennum, Emmanuel Mignot, Helge Bjarup Dissing Sorensen

JOURNAL: Sleep

STATUS: Under review

FULL CITATION: A. N. Olesen, P. Jennum, E. Mignot and H. B. D. Sorensen, "Automatic sleep stage classification with deep residual networks in a mixed-cohort setting," *Sleep*, 2020, *under review*

TITLE PAGE

Title

Automatic sleep stage classification with deep residual networks in a mixed-cohort setting

Authors and affiliations

Alexander Neergaard Olesen[1,2,3], Poul Jennum[3*], Emmanuel Mignot[2*], Helge Bjarup Dissing Sørensen[1*],

[1]Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark

[2]Stanford Center for Sleep Sciences and Medicine, Stanford University, Palo Alto, CA, USA

[3]Danish Center for Sleep Medicine, Department of Clinical Neurophysiology, Rigshospitalet, Glostrup, Denmark

*These authors have contributed equally

Corresponding author

Alexander Neergaard Olesen, aneol@dtu.dk

Department of Health Technology, Technical University of Denmark

Ørsteds Plads, building 349, room 010

2800 Kgs. Lyngby, Denmark

Institution where work was performed

Stanford Center for Sleep Sciences and Medicine, Stanford University

Department of Health Technology, Technical University of Denmark

ABSTRACT

Study Objectives: Sleep stage scoring is performed manually by sleep experts and is prone to subjective interpretation of scoring rules with low intra- and interscorer reliability. Many automatic systems rely on few small-scale databases for developing models, and generalizability to new datasets is thus unknown. We investigated a novel deep neural network to assess the generalizability of several large-scale cohorts.

Methods: A deep neural network model was developed using 15.684 polysomnography studies from five different cohorts. We applied four different scenarios: 1) impact of varying time-scales in the model; 2) performance of a single cohort on other cohorts of smaller, greater or equal size relative to the performance of other cohorts on a single cohort; 3) varying the fraction of mixed-cohort training data compared to using single-origin data; and 4) comparing models trained on combinations of data from 2, 3, and 4 cohorts.

Results: Overall classification accuracy improved with increasing fractions of training data (0.25%: 0.782 ± 0.097, 95% CI [0.777 – 0.787]; 100%: 0.869 ± 0.064, 95% CI [0.864 – 0.872]), and with increasing number of data sources (2: 0.788 ± 0.102, 95% CI [0.787 – 0.790]; 3: 0.808 ± 0.092, 95% CI [0.807 – 0.810]; 4: 0.821 ± 0.085, 95% CI [0.819 – 0.823]). Different cohorts show varying levels of generalization to other cohorts.

Conclusions: Automatic sleep stage scoring systems based on deep learning algorithms should consider as much data as possible from as many sources available to ensure proper generalization. Public datasets for benchmarking should be made available for future research.

Keywords

Automatic sleep stage classification, computational sleep science, machine learning, deep learning

STATEMENT OF SIGNIFICANCE

Manual annotation of polysomnography studies is subject to human bias with multiple studies showing variations in how sleep experts score sleep. Most research in automatic sleep stage classification models use small-scale data from a single origin, and it is unknown how these models generalize to new data. We developed an algorithm for automatic scoring of sleep stages using raw polysomnography data and obtain state-of-the-art classification performance on a large number of test subjects. Our algorithm was tested under different conditions to compare generalizability. We found that using data from many different sources improves classification performance, and that models trained on single-origin data generalize inconsistently to new data. Future researchers should take multiple datasets into account when developing sleep scoring models.

INTRODUCTION

Sleep staging is important to the analysis of human sleep with about 845,000 sleep studies performed in 2014 in the US alone[1]. Briefly, a standard clinical sleep study consists of a full-night polysomnography (PSG) comprising electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), electrocardiography (ECG), thoraco-abdominal inductance plethysmography, oronasal thermal flow, nasal pressure, and blood saturation recordings. These studies are then evaluated by experts for the presence of events of clinical relevance, as determined by standards created by the American Academy of Sleep Medicine (AASM), such as the number of blood oxygen desaturations, micro-arousals, leg movements, periods of cessated breathing, etc. Furthermore, the overall sleep architecture is captured in a hypnogram conducted by labeling every 30 s of PSG data into one of five stages of sleep: wakefulness (W), rapid eye movement (REM) sleep, and non-REM stage 1, 2, and 3 (N1, N2, N3). The latter three stages are distinguished by distinct EEG amplitude and frequency distributions, the presence of specific EEG micro-events and arousability differences reflecting sleep depth. Sleep stage labeling is summarized in key metrics, such as the percentage of total sleep time (TST) spent in any of the five stages (%W, or wake after sleep onset, WASO; %REM; %N1; %N2; %N3), and visually in the form of a hypnogram, which shows temporal progression of sleep stages across the night. Current clinical practice (gold standard) of sleep study analysis is manual scoring and annotation of sleep stages and sleep events based on guidelines from the AASM[2]. These guidelines, based on observations made in healthy young males almost 70 years ago are problematic for several reasons: a) technicians will never score the same data the exact same way as another technician, or even the same way twice[3–7]; b) normal sleep from healthy young males may not reflect sleep patterns of patients referred to sleep clinics; and c) the 30 s epoch rule is arbitrary and was based on physical limitations of recording equipment when PSGs were recorded on paper.

Automatic sleep stage classification has not yet seen wide-spread adoption in clinical practice despite ongoing research demonstrating feasibility and industrial interests[8]. A major issue has been a lack of available data for designing and training models. The publicly available PhysioNet Sleep-EDF and the expanded version[9,10] has been used extensively for training both shallow and deep learning-based machine learning models[11–13], but given its small sample size and homogeneity (most papers use the same healthy 20 subjects), it is questionable how well models derived from this data generalize to unseen data, even if high classification performance is often reported[8]. Other databases which have been extensively used include the St. Vincent's University Hospital and University College Dublin Sleep Apnea Database ($n = 25$)[9,14], and the Montreal Archive of Sleep Studies (MASS, $n = 200$)[13,15–19]. The argument for using deep learning-based models to classify high-dimensional electrophysiological data, e.g. PSGs, into discrete outcomes such as sleep stages is compelling, because of their ability to capture variability in the underlying, highly complex, data representations, that might be missed by machine learning methods relying on manual feature engineering. In the image, speech, and natural language processing domains, the success of deep learning models using untransformed data have been unsurpassed in the last decade, thanks largely due to the availability of ever-increasing amounts of compute resources and more significantly very large, robust and diverse datasets[20].

Recently deep learning models for automatic sleep stage classification have been developed and validated using two or more databases or cohorts[21–23], or using a single large volume cohort[22,24,25]. The assumption has been that by incorporating multiple sources of variance in the dataset used for training (e.g. from multiple technicians, sites, recording setups, equipment, etc.), final models will be better at generalizing to new, unseen data. However, no study to date has investigated multiple, large-scale cohorts for automatic sleep stage classification, or how different cohorts generalize to one another.

In this work, we describe a deep learning-based sleep stage classification algorithm trained and validated on raw PSG data from multiple, large-scale cohorts for a total of 15,684 studies, that outputs a probability distribution over all sleep stages at a given time resolution. Considering the amount of data available, our aim was to evaluate: 1) how well does performance of individual cohorts generalize to others; 2) how much data is needed for accurate sleep staging; 3) how many cohorts are necessary for that same goal; and 4) which is better, more data, or more diverse data. To our knowledge, this is one of the largest, if not the largest, study on automatic sleep stage classification in terms of PSG volume and diversity.

METHODS

Cohort descriptions

To investigate and conclude on generalizability of any machine learning or sleep stage classification model, multiple heterogenous datasets must be used for training, validation and testing purposes. In this work, we collected datasets from five different sources, each dataset containing a diverse collection of subjects presenting with multiple disease phenotypes. Details of the separate cohorts are shown in Table 1 along with reported *p*-values highlighting cohort differences. Each cohort was split into a training, validation and testing *subset* in proportions of 87.5%, 2.5% and 10%, respectively, using random sampling without replacement among unique subjects, so that no subject is shared between subsets. With these percentages, we maximize the number of PSGs available for training, while still reserving enough PSGs for validation and testing. Collecting all the separate subsets across cohorts forms a training, validation, and testing *partition*, containing the respective subsets from all five cohorts.

Institute of Systems and Robotics, University of Coimbra Sleep Cohort (ISRUC)

This cohort contains 126 recordings from 118 unique subjects recorded at the Sleep Medicine Centre of the Hospital of Coimbra University, Portugal, in the period 2009–2013[26]. The cohort comprises three subgroups: subgroup I contains 100 PSGs of subjects with diagnosed sleep disorders, generally sleep apnea; subgroup II contains 16 recordings of eight subjects most of which are also diagnosed with sleep apnea; and subgroup III contains recordings from 10 subjects with no diagnosed sleep disorders. All PSGs were recorded with the same recording hardware and software and each was scored by two technicians for sleep stages and sleep events according to the AASM guidelines. ISRUC-Sleep is a freely accessible resource and all data and PSG files can be located at https://sleeptight.isr.uc.pt/ISRUC_Sleep/.

The MrOS Sleep Study (MrOS)

The MrOS sleep study is part of the larger Osteoporotic Fractures in Men Study, which aims to understand the relationships between sleep disorders, fractures, and vascular diseases in community-dwelling men[27–29]. It consists of 2,907 in-home PSG recordings with an additional 1,026 follow-up PSG studies from subjects recruited from six different clinical centers in the USA. Each recording was annotated by an expert technician according to Rechtschaffen and Kales (R&K) criteria for sleep staging[30]. For compatibility with AASM guidelines, we combined stages labeled S3 and S4 into N3. All data were accessed from the National Sleep Research Resource (NSRR) repository[31,32].

The Sleep Heart Health Study (SHHS)

The SHHS is a large, multi-center study on cardiovascular outcomes related to sleep disorders with a specific focus on sleep-disordered breathing[33,34]. The cohort consists of 6,441 subjects above 40 years old recruited between 1995 and 1998 undergoing in-home PSG (SHHS Visit 1) with subsequent follow-up PSG between 2001 and 2003 in 3,295 subjects (SHHS Visit 2). PSG recordings were annotated for sleep stages by trained and certified technicians according to R&K rules. From the original cohort we extracted 5,793 PSGs and annotations from Visit 1, and 2,651 from Visit 2, and aggregated S3 and S4 stages into N3 similar to MrOS. All data were accessed from NSRR repository.

Wisconsin Sleep Cohort (WSC)

WSC is a population-based study of sleep-disordered breathing in government workers in Wisconsin, USA that was initiated in 1988[35,36]. In this work, we used 2412 PSGs from 1091 unique subjects in the WSC sample scored by expert technicians according to R&K rules with subsequent merging of S3 and S4 into N3.

Stanford Sleep Cohort (SSC)

PSGs from this cohort originate from patients referred for sleep disorders evaluation and recorded at the Stanford Sleep Clinic since 1999. The specific sample used in this study represents a small subset (n = 772) of the whole cohort, which was selected and described in detail in previous studies[37,38] scored according to R&K or AASM guidelines according to prevailing standard at the time of evaluation.

Signal pre-processing pipeline

Electrophysiological signals corresponding to the minimum acceptable montage for sleep staging available across all cohorts were extracted for each PSG. These included a central EEG (either C3 or C4 referenced to the contra-lateral mastoid), left and right EOG referenced to the contra-lateral mastoid, and a single submentalis EMG. The choice between C3 and C4 was determined based on the lowest total signal energy across the entire duration of the PSG to avoid excessive signal popping. Other methods to determine appropriate channels include algorithms based on shortest Mahalanobis distance to an already determined reference distribution[21], but was not investigated in this study. All signals were resampled to f_s = 128 Hz using a polyphase filtering procedure irrespective of original sampling frequency; and subsequently filtered using a zero-phase approach with 4th order Butterworth IIR filters (0.5 to 35 Hz band pass for EEG and EOG; 10 Hz high pass for EMG) in accordance with AASM filter specifications[2]. Each signal was normalized to zero mean and unit variance to accommodate differences in recording equipment and baselines; and to compress the dynamic range into something easily trainable for the neural network architecture. We denote by C the number of input signals supplied to the neural network, where in this case $C = 4$.

Machine learning problem

We designate by $\mathcal{X} \in \mathbb{R}^{C\times T}$ the set of 30 s input data segments with *C* input channels and segment length *T*, and the corresponding classifications by $\mathcal{Y} = \{y \in \mathbb{R}_{+}^{K} \,|\, \sum_i y_i = 1\}$, where *K* = 5 corresponds to the five sleep stages. Thus, y is a probability simplex, which maps to the ordered set $\mathcal{S} = \{\mathrm{W}, \mathrm{N1}, \mathrm{N2}, \mathrm{N3}, \mathrm{REM}\}$ by the argmax function such that $\operatorname{argmax} y : \mathcal{Y} \rightarrow \mathcal{S}$. Furthermore, as we are potentially interested in classifying multiple sleep stages at once, we extend the problem of classifying a single sleep stage given $x \in \mathcal{X}$ to a sequence-to-sequence problem, in which we desire to learn a differentiable function representation Φ, that maps a sequence of 30 s epochs $\mathbf{x} \in \mathbb{R}^{C\times\alpha T}$ to their corresponding label probabilities $\mathbf{y} \in \mathbb{R}^{K\times\alpha}$, where α is a parameter that controls the sequence length. If e.g. $\alpha = 8$, the sequence x contains 4 min of successive PSG data described by 8 epochs of length 30 seconds. Furthermore, we denote by $[\![a, b]\!]$ the set of integers from a to b, i.e. $[\![a, b]\!] \equiv \{n \in \mathbb{N} | a \leq n \leq b\}$, and by $[\![N]\!]$ the shorthand form of $[\![1, N]\!]$.

Network architecture

As the representation of Φ, we adapted and extended a previously published neural network architecture for automatic sleep stage classification, which was based on a variant of the ResNet-50 architecture commonly used for two-dimensional image classification tasks, but adapted and re-trained from scratch for the specific use-case of one-dimensional, time-dependent signals in the PSG[24]. This network has the advantage that it does not require any manual feature engineering and extraction compared to previous state of the art sleep stage classification models[21]. An overview of the proposed network architecture is provided graphically in Figure 1 and Table 2. Briefly, the architecture consists of four modules:

1) an initial mixing module $\varphi_{\text{mix}} : \mathbb{R}^{1\times C\times T} \rightarrow \mathbb{R}^{C\times 1\times T}$
2) a feature extraction module $\varphi_{\text{feat}} : \mathbb{R}^{C\times 1\times T} \rightarrow \mathbb{R}^{f_0 2^{R+1}\times 1\times T/2^R}$
3) a temporal processing module $\varphi_{\text{temp}} : \mathbb{R}^{f_0 2^{R+1}\times 1\times T/2^R} \rightarrow \mathbb{R}^{2n_h\times T/2^R}$, and
4) a classification module $\varphi_{\text{clf}} : \mathbb{R}^{2n_h\times T/2^R} \rightarrow \mathbb{R}^{K\times T/2^R}$.

Thus, we obtain a differentiable representation of the function $\Phi : \mathbb{R}^{C\times T} \rightarrow \mathbb{R}^{K\times T/2^R}$ as

$$\Phi(\mathbf{x}) = \varphi_{\text{clf}}\left(\varphi_{\text{temp}}\left(\varphi_{\text{feat}}(\varphi_{\text{mix}}(\mathbf{x}))\right)\right).$$

The output of this function is the matrix $\mathbf{y} \in \mathbb{R}^{K\times T/2^R}$ containing sleep stage probabilities in the sequence of PSG data evaluated every second.

Mixing module

The raw input data is input to this module, which encourages non-linear channel mixing similar to what has been proposed in recent literature[16,39–41]. The module is realized using a single 2D convolutional operation outputting C feature maps computed using single-strided $C \times 1$ kernels followed by rectified linear unit (ReLU) activations.

Feature extraction (residual network) module

This is comprised of a succession of R residual blocks (see Figure 1), which are responsible for the bulk feature extraction from the channel-mixed data. Each residual block is realized using bottlenecks of first a 1×1 convolution to reduce the number of feature maps, then a 1×3 convolution and lastly a 1×1 convolution to finally increase the number of feature maps. Each convolution operation was followed by a batch normalization[42] and ReLU activation except after the last convolutional layer, where shortcut projections are added before the activation[43]. This type of block structure enables the design and training of very deep networks without the risk of vanishing gradients due to the projection shortcuts[44].

Temporal processing module

This module is realized by a bidirectional gated recurrent unit (GRU)[45] in order to accommodate temporal dependencies in the PSG. The GRU runs through the temporal dimension of the output from φ_{feat} of $T/2^R$ time steps each containing $f_0 2^{R+1}$ features and outputs n_h features in each direction for each time step. By running both forward and backward, we can accommodate that technicians base their scoring on looking backwards as well as ahead in time in each time segment (typically 30 s).

Classification module

The final module in the architecture performs actual classification based on the forward and backward features for each time step outputted from φ_{temp}. It is realized by a single convolutional operation with a subsequent softmax activation to compute a probability distribution over the K sleep stage classes, such that the probability of sleep stage i at time step n is given by $y_i^{(n)} = \frac{\exp{(a_i)}}{\sum_k \exp{(a_k)}}$, where $a_i \in \boldsymbol{a}$ is the activation of the last layer in the network and $k = [\![K]\!]$.

Loss function

The network was trained end-to-end with respect to a loss function, that takes the output probabilities from the network $\mathbf{y} = \Phi(\mathbf{x})$ and calculates the loss as

$$\mathcal{L}(\mathbf{y}) = -\sum_{n=1}^{30/\tau} \sum_{k=1}^{K} t_k{}^{(n)} \log\left(y'^{(n)}_{k}\right), \tag{1}$$

$$y'^{(n)}_k = \frac{1}{\tau} \sum_{i=\tau(j-1)+1}^{\tau n} y^{(i)}_k, \tag{2}$$

which is the cross-entropy between successive time-averaged classifications (parameterized by the number of successive one-second predictions τ), and the ground truth labels t broadcasted to $30/\tau$ labels per 30 s segment. This way, we can acquire predictions every second, that can be combined in time at intervals given by τ.

Experimental setups

We set up three different experiments in this study.

A) We wished to investigate the effect of increasing the complexity of the recurrent module by varying the number of units n_h in the module φ_{temp} in the space $n_h = 2^k$, $k \in [\![6,11]\!]$. We hypothesize that there exists a sweet-spot in the number of hidden units that balances computational complexity with classification performance, i.e. classifying a sequence of sleep stage labels given a corresponding sequence of outputs from φ_{feat}. The results of this experiment were furthermore used to determine parameters for models in subsequent experiments.

B) Since we have several cohorts at our disposition of both clinical and research origin, we can investigate the compatibility and inherent generalizability of the different cohorts in two ways: 1) we set aside a single cohort for testing, while we train the models on the remaining four (leave-one-cohort-out, LOCO training); and 2) we train on a single cohort, while we set aside the remaining four for testing (leave-one-cohort-in, LOCI training).

C) Generalizability can also be investigated in another way, which can answer the question of how many data sources is necessary. We trained models with all possible 2-, 3-, and 4-combinations of cohorts, i.e. one run trained on ISRUC and MrOS training data, another run with ISRUC and SHHS train data, a third with ISRUC and SSC, etc., with all runs subjected to subsequent evaluation on the test partition.

D) Previous studies have already investigated the performance of automatic sleep staging algorithms using shallow machine learning models. At the time of writing however, none have investigated the effect of available training data for deep learning models at this magnitude (up to tens of thousands). We therefore trained models on 0.25%, 0.5%, 1%, 5%, 10%, 25%, 50%, 75% and 100% of the data available for training. Specifically, some of these fractions of the total number of PSGs correspond roughly to the number of PSGs in the training partitions in each cohort, allowing for direct comparisons between training a model with mixed- and single-cohort training data.

Common for all experiments were the default parameter values $C = 4$, $f_s = 128$ Hz, $T = \tau f_s$, $K = 5$, $R = 7$, and $f_0 = 4$ for the number of input channels, sampling frequency, the sequence length, the number of sleep stages, the number of consecutive residual blocks, and the base filter kernel size, respectively. All models were trained for 50 epochs (passes through the training partition) and the model with the highest Cohen's kappa value on the validation partition was subsequently selected for testing. All models were trained end-to-end with backpropagation using the Adam optimizer[46] with a learning rate of 10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$ to minimize the loss function specified by Eq. (1) and Eq. (2). All network weights and bias terms were initialized using the uniform Glorot initialization scheme[47].

Performance metrics and model evaluation

For each experiment we evaluated model performance using the overall accuracy (Acc) and Cohen's kappa (κ) in order to into account the possibility of chance agreement between the model gold standard. Given a confusion matrix $\mathbf{C}$ with element c_{ij} being the number of epochs belonging to sleep stage i but classified to be in sleep stage j, we define the overall accuracy for a given model as

$$\text{Acc} = \frac{\sum_{i=j} c_{ij}}{\sum_{i,j} c_{ij}}$$

i.e. the sum of the trace of $\mathbf{C}$ divided by the total count. The Cohen's kappa metric is defined as

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where $p_o = \text{Acc}$ is the observed agreement (i.e. accuracy) and p_e is the expected chance agreement, which can be reformulated in terms of the outer product between the row and column sums (class-specific recall and precision) of $\mathbf{C}$.

Data and source code availability

All model training and testing code was implemented in PyTorch v. 1.2[48]. Model performances were assessed using custom Python scripts using scikit-learn[49]. Source code and pre-trained models will be made available at https://github.com/neergaard/deep-sleep-pytorch.git and https://github.com/Stanford-STAGES/deep-sleep-pytorch.git upon publication of this paper. Data from ISRUC are publicly available at https://sleeptight.isr.uc.pt/ISRUC_Sleep/, while access to data from MrOS and SHHS can be requested from the NSRR. Anonymized PSG data from SSC including selected demographic data are available at https://stanfordmedicine.app.box.com/s/r9e92ygq0erf7hn5re6j51aaggf50jly.

RESULTS

In this section we report on the results of the three experiments described in the **Experimental setups** section.

Temporal context impact on model performance

In Figure 2 we show how the model performance depends on the temporal context and complexity of the temporal processing module, when evaluating the model on the validation partition. Results are further detailed in Table S1. Specifically, we observe a drastic change in Cohen's kappa just by introducing a simple recurrent unit into the network as shown in Figure 2a, where Cohen's kappa increases from 0.645 ± 0.126 (95% CI: $[0.633 - 0.657]$) at $n_h = 0$ to 0.720 ± 0.120 (95% CI: $[0.709 - 0.731]$) at $n_h = 64$. We did not observe any major changes when increasing the number of hidden units beyond 64, although we did see a maximum Cohen's kappa of 0.734 ± 0.111 (95% CI: $[0.723 - 0.744]$) at $n_h = 1024$, which is shown in the inset in Figure 2a. We observed a general increase in Cohen's kappa when classifying longer sequences than 2 min (0.726 ± 0.114, 95% CI: $[0.715 - 0.737]$), but did not see any major differences when classifying over more than 3 min sequences (0.733 ± 0.123, 95% CI: $[0.721 - 0.744]$). Subsequent models were fixed with $n_h = 1024$ corresponding to a sequence length of 5 min.

Model classifications converge to 30 s predictions given sufficient training data

Furthermore, we analyzed the classification performance of the model given a specific sequence length by looking at the average prediction accuracy across all 5 min sequences in all subject PSGs in the test partition, similar to what Brink-Kjaer et al. has shown previously[50]. In Figure 2c, we show how the average classification accuracy in a 5 min sequence both depends on the amount of data and the frequency of evaluating the model output, i.e. every 1 s or across 30 s. The average classification accuracy was found to be slightly lower in the beginning of each 5 min sequence (see Figure 2c), both when training a model with less (500 training subjects) and more (75% of total training subjects). Interestingly, when training with less data, we also observed a lower accuracy in the beginning and end of each 30 s segment relative to the accuracy in the middle section, which was not the case when training with more data.

Choice of cohort impacts classification performance on test set

In Figure 3 we show how training on different cohorts yield differing results in subsequent testing performance, here expressed in heatmaps as both overall accuracy (Figure 3a), and Cohen's kappa (Figure 3b) averaged across all $N = 1{,}584$ subject PSGs in the test partition. The first two columns show the performance on the cohort on the x-axis, when training on the specific cohort on the y-axis. Since the training subset in ISRUC is small compared to the other cohorts, we trained the model in the left-most column with weight decay of 10^{-4} to compensate for the risk of overfitting, however, by comparing the left and middle columns, we did not observe any specific gain in classification performance by doing so. The right-most column shows the test performance for each cohort, when excluding that cohort from training. We observe a significant spread in classification accuracy across the different cohorts with prediction on ISRUC being poorest, while prediction on MrOS data being best. Further details can be found in Table S2.

More data is good, diverse data is better

We observed a general increase in classification performance both in terms of overall accuracy and Cohen's kappa, when including more data in the model training phase in both the mixed- and single-cohort setting (Figure 4a, Table S3). Classification performance was consistently lower in the single-cohort setting compared to the corresponding mixed-cohort setting. Interestingly, we found that training a model with just 0.25% of mixed-cohort training data still achieved an acceptable accuracy comparable to training a model with only SHHS data, while using all available training data increased that performance by almost 10 percentage points. Furthermore, we observed that the model trained with 100% of the training partition reached a state-of-the-art level of performance with an overall accuracy of 0.869 ± 0.064 (95% CI: $[0.865 - 0.872]$) and Cohen's kappa of 0.799 ± 0.098 (95% CI: $[0.794 - 0.804]$) (Table S3). The model furthermore performs well with respect to classifying individual sleep stages as shown in the confusion matrix in Figure 4b. However, the model still has difficulties classifying and distinguishing between certain sleep stages, especially between N2, N1, and N3; and W, N2, and N1.

Increasing the number of data sources improves classification performance

On average, we saw an increase in overall accuracy, when increasing the number of cohorts from 2 to 4 using 500 PSGs in each configuration, see Figure 5 and Table S4. Specifically, we found that the average overall accuracy increased from 0.788 ± 0.102 (95% CI: $[0.787 - 0.790]$) in the 2-cohort configuration to 0.808 ± 0.092 (95% CI: $[0.807 - 0.810]$) and 0.821 ± 0.085 (95% CI: $[0.819 - 0.823]$) in the 4-cohort configuration.

DISCUSSION

In this work, we present an end-to-end deep learning-based model for fully automatic micro- and macro-sleep stage classification. Using all of the available data sources for training our model, we reached an overall accuracy on test partition of 0.869 ± 0.064 (95% CI: $[0.865 - 0.872]$), and a Cohen's kappa of 0.799 ± 0.098 (95% CI: $[0.794 - 0.804]$), which is in the very high end of the substantial agreement category for observer agreement[51]. We found that individual cohorts exhibit major differences in overall accuracy and Cohen's kappa when subjected to both training and testing conditions and specifically, we found that average performance on the test partition in the LOCI configurations varied significantly from 0.676 ± 0.124 (95% CI: $[0.670 - 0.682]$) when training on ISRUC, to 0.837 ± 0.084 (95% CI: $[0.833 - 0.841]$) when training on SHHS. Each individual cohort also showed large deviations in predictive performance when tested on the other cohorts. For example, when conditioned on SHHS data, the lowest average accuracy was 0.721 on SSC test data compared to the highest at 0.872 on SHHS test data, while conditioning on SSC training data, the lowest average accuracy was 0.704 on ISRUC test data compared to 0.824 on WSC test data. Classification performance was generally higher on the test set when using the LOCO configuration, except for SHHS (higher in LOCI) and SSC (no difference). We also found that having data from multiple sources always resulted in better-performing models compared to training on single cohorts. Increasing the number of data sources increased classification performance, although this was non-significant. In the design of the model, we observed that model performance was enhanced by the addition of the recurrent module (bGRU), a phenomenon likely reflecting the fact that sleep stage scoring at a specific time in one subject can be influenced by signal content (frequency, amplitude, presence of micro-events) at later time steps. However, the complexity of the module given by the number of hidden units did not affect performance. In all our experiments, we also evaluated the performance of the model every 1 s compared to the performance evaluated every 30 s and found them to be similar, which indicates the model is stable in classification in periods corresponding to an epoch of data.

Only a handful of studies have previously reported results when using multiple cohorts[21–23]. Some authors have reported a drop from 81.9% to 77.7% when training on the Massachusetts General Hospital cohort (MGH) and testing on MGH and SHHS, respectively[22], while others have shown significant drops from 89.8% to 81.4% and 72.1% on two separate hold-out sets from Singapore and USA[23]. We also observed similar trends in our LOCI and LOCO experiments, where excluding the training subset of a cohort from the training partition resulted in a significant drop in performance on the respective test subset from that cohort. A benefit of our LOCI and LOCO experiments is the possibility for direct benchmarking against previous publications using specific cohorts in their experiments. For example, we obtain an accuracy of 0.805 in the LOCO-SHHS training-testing case compared to 0.777 previously reported by Biswal et al.[22], both of which reflect classification performance when SHHS had not been used for training; and an accuracy of 0.865 in the LOCI-WSC case compared to 0.841 reported by Olesen et al. [24], where both have been using a subset of WSC for training the model. Interestingly, we obtained the same level of performance on the SHHS data in our LOCI experiment as reported by Sors et al. (87% accuracy, 81% Cohen's kappa) even though they only used single-EEG for their experiments[52]. Other works that have investigated single- vs. multi-channel models for automatic sleep stage classification have found that models generally benefit from having more channels available for training[16,18,22]. It may be that some cohorts share different characteristics that makes them more suitable for single- or multi-channel models, but this is speculative and would need to be verified in subsequent studies.

Our study is not without limitations. We only optimized our network architecture with respect to the temporal processing module and therefore cannot assess what impact different design choices for the other modules would have had on final performance. For example, the EMG signal has different statistical properties and spectral content, and separate, parallel architectures for EMG and EEG/EOG feature extraction may be warranted, as proposed by others[16,21]. Other studies have however shown equal performance in large cohorts using a similar channel mixing approach as proposed here[24]. Another limitation is found in our training runs, as we did not consider balancing our data with respect to the proportion of sleep stages, which may or may not have had impact on overall performance. It is well established that there is significant variation in scoring and validation of N1/REM and N2/N3[3,5,7], which challenges the training for any classification algorithm. Some researchers have experimented balancing the cost of misclassifying sleep stages by weighting them by their inverse frequency of occurrence and found no significant improvement[24,52], while others have experimented with balancing the sleep stage frequencies in each batch of data input to the neural network model[16], but more rigorous research in resampling or over/under-sampling techniques is warranted in this regard. We ultimately decided against experimenting with balancing our sleep stages in each batch, as we prioritized flexibility with regards to the length of input sequences fed to the network. All our models ran through at least 50 epochs of training (passes through the training partition), which might have induced a bias in the configurations with larger cohorts. For example, one pass through the training partition in the LOCI-ISRUC case corresponds to much less data than one pass through the LOCI-SHHS case. However, since we selected the best performing model based on Cohen's kappa across all 50 epochs, we have allowed for more effective training in cases with less available training data. We observed that models using less data in the training partition generally had to run for longer time (i.e. more epochs) before converging.

In future studies on automatic sleep stage classification algorithms, we strongly recommend researchers to test and report results on not just hold-out test partitions, but also on cohorts completely unseen by the model both during training and testing/validation. Our experiments indicate that even though good performance can be achieved on hold-out data using a single cohort, this does not necessarily translate into good generalization performance. Such approach requires availability of many publicly available, high-quality, well-documented databases with easily accessible PSG data, associated annotations and related patient information. In this regard, websites such as the NSRR, which contains several large databases with clinical data as well as PSG and annotation data in a standardized format[31,32], are an invaluable resource for researchers. We also propose that the sleep science community establishes a common reference dataset on which researchers in machine learning can benchmark their models, similar to what the computer vision and general machine learning community has done with the ImageNet Large Scale Visual Recognition Challenge (ILSVRC)[53], an annual competition in which researchers submit their models to test in various competitions.

In summary, we have developed an automatic sleep stage classification algorithm based on deep learning, that can accurately classify sleep stages at a flexible resolution with a state-of-the-art classification performance of 87% accuracy on a test set of 1,584 PSGs. We trained and tested our model using five cohorts with varying numbers of PSGs covering multiple phenotypes with specific focus on how well cohorts can generalize to each other. We found that different cohorts generalize very differently both in intra- and inter-cohort settings (LOCI vs. LOCO experiments). Furthermore, we also found that having more data sources significantly improve classification performance and generalizability to the extent that even just a small number of training PSGs can reach high classification performance by including many different sources. To our knowledge, this is one of the largest, if not the largest, study on automatic sleep stage classification in terms of PSG volume, diversity, and performance.

ACKNOWLEDGMENTS

Some of the computing for this project was performed on the Sherlock cluster. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results.

The authors would also like to thank the National Sleep Research Resource (https://www.sleepdata.org) team for their work in collecting, organizing and making available some of the PSG data used in this study.

The authors would also like to thank Julie Anja Engelhard Christensen, PhD, for her editorial work on this paper.

DISCLOSURE STATEMENT

A. N. Olesen has received funding from The Klarman Family Foundation; Technical University of Denmark; University of Copenhagen; Reinholdt W. Jorck og Hustrus Fonden; Otto Mønsteds Fond; Stibofonden; Knud Højgaards Fond; Augustinus Fond; and Vera og Carl Johan Michaelsens Fond. E. Mignot has received partial funding by the Klarman Family Foundation, has received research support from Jazz Pharmaceuticals, is a consultant for Rhythm (a sleep consumer product company) and Alairion (a sleep apnea pharmacology company), and is on the speakers' bureau for Vox Media.

REFERENCES

1. Chiao W, Durr ML. Trends in sleep studies performed for Medicare beneficiaries. *Laryngoscope*. 2017;127(12):2891-2896. doi:10.1002/lary.26736
2. Berry RB, Albertario CL, Harding SM, et al. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications.* 2.5. Darien, Il: American Academy of Sleep Medicine; 2018.
3. Younes M, Raneri J, Hanly P. Staging sleep in polysomnograms: Analysis of inter-scorer variability. *J Clin Sleep Med*. 2016;12(6):885-894. doi:10.5664/jcsm.5894
4. Younes M. The case for using digital EEG analysis in clinical sleep medicine. *Sleep Sci Pract*. 2017;1(2). doi:10.1186/s41606-016-0005-0
5. Younes M, Kuna ST, Pack AI, et al. Reliability of the American Academy of Sleep Medicine Rules for Assessing Sleep Depth in Clinical Practice. *J Clin Sleep Med*. 2018;14(02):205-213. doi:10.5664/jcsm.6934
6. Rosenberg RS, Van Hout S. The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring. *J Clin Sleep Med*. 2013;9(1):81-87. doi:10.5664/jcsm.2350
7. Norman RG, Pal I, Stewart C, Walsleben JA, Rapoport DM. Interobserver Agreement Among Sleep Scorers From Different Centers in a Large Dataset. *Sleep*. 2000;23(7):1-8. doi:10.1093/sleep/23.7.1e
8. Fiorillo L, Puiatti A, Papandrea M, et al. Automated sleep scoring: A review of the latest approaches. *Sleep Med Rev*. 2019;48:101204. doi:10.1016/j.smrv.2019.07.007
9. Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*. 2000;101(23):e215-e220. doi:10.1161/01.CIR.101.23.e215
10. Kemp B, Zwinderman AH, Tuk B, Kamphuisen HAC, Oberyé JJL. Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG. *IEEE Trans Biomed Eng*. 2000;47(9):1185-1194. doi:10.1109/10.867928
11. Vilamala A, Madsen KH, Hansen LK. Deep convolutional neural networks for interpretable analysis of EEG sleep stage scoring. In: *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*. Tokyo, Japan: IEEE; 2017:1-6. doi:10.1109/MLSP.2017.8168133
12. Phan H, Andreotti F, Cooray N, Chen OY, Vos M De. Automatic Sleep Stage Classification Using Single-Channel EEG: Learning Sequential Features with Attention-Based Recurrent Neural Networks. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE; 2018:1452-1455. doi:10.1109/EMBC.2018.8512480
13. Supratak A, Dong H, Wu C, Guo Y. DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG. *IEEE Trans Neural Syst Rehabil Eng*. 2017;25(11):1998-2008. doi:10.1109/TNSRE.2017.2721116
14. Şen B, Peker M, Çavuşoğlu A, Çelebi F V. A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms. *J Med Syst*. 2014;38(3). doi:10.1007/s10916-014-0018-0
15. O'Reilly C, Gosselin N, Carrier J, Nielsen T. Montreal archive of sleep studies: An open-access resource for instrument benchmarking and exploratory research. *J Sleep Res*. 2014;23(6):628-635. doi:10.1111/jsr.12169
16. Chambon S, Galtier MN, Arnal PJ, Wainrib G, Gramfort A. A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series. *IEEE Trans Neural Syst Rehabil Eng*. 2018;26(4):758-769. doi:10.1109/TNSRE.2018.2813138
17. Andreotti F, Phan H, Cooray N, Lo C, Hu MTM, De Vos M. Multichannel Sleep Stage Classification and Transfer Learning using Convolutional Neural Networks. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE; 2018:171-174. doi:10.1109/EMBC.2018.8512214
18. Phan H, Andreotti F, Cooray N, Chen OY, De Vos M. Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification. *IEEE Trans Biomed Eng*. 2019;66(5):1285-1296. doi:10.1109/TBME.2018.2872652

19. Phan H, Andreotti F, Cooray N, Chen OY, De Vos M. 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. doi:10.1109/TNSRE.2019.2896659

20. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539

21. Stephansen JB, Olesen AN, Olsen M, et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nat Commun*. 2018;9(1):5229. doi:10.1038/s41467-018-07229-3

22. Biswal S, Sun H, Goparaju B, Westover MB, Sun J, Bianchi MT. Expert-level sleep scoring with deep neural networks. *J Am Med Informatics Assoc*. 2018;25(12):1643-1650. doi:10.1093/jamia/ocy131

23. Patanaik A, Ong JL, Gooley JJ, Ancoli-Israel S, Chee MWL. An end-to-end framework for real-time automatic sleep stage classification. *Sleep*. 2018;41(5):1-11. doi:10.1093/sleep/zsy041

24. Olesen AN, Jennum P, Peppard P, Mignot E, Sorensen HBD. Deep residual networks for automatic sleep stage classification of raw polysomnographic waveforms. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE; 2018:1-4. doi:10.1109/EMBC.2018.8513080

25. Biswal S, Kulas J, Sun H, et al. SLEEPNET: Automated Sleep Staging System via Deep Learning. July 2017:1-17. http://arxiv.org/abs/1707.08262.

26. Khalighi S, Sousa T, Santos JM, Nunes U. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. *Comput Methods Programs Biomed*. 2016;124:180-192. doi:10.1016/j.cmpb.2015.10.013

27. Blank JB, Cawthon PM, Carrion-Petersen M Lou, et al. Overview of recruitment for the osteoporotic fractures in men study (MrOS). *Contemp Clin Trials*. 2005;26(5):557-568. doi:10.1016/j.cct.2005.05.005

28. Orwoll E, Blank JB, Barrett-Connor E, et al. Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study — A large observational study of the determinants of fracture in older men. *Contemp Clin Trials*. 2005;26(5):569-585. doi:10.1016/j.cct.2005.05.006

29. Blackwell T, Yaffe K, Ancoli-Israel S, et al. Associations Between Sleep Architecture and Sleep-Disordered Breathing and Cognition in Older Community-Dwelling Men: The Osteoporotic Fractures in Men Sleep Study. *J Am Geriatr Soc*. 2011;59(12):2217-2225. doi:10.1111/j.1532-5415.2011.03731.x

30. Rechtschaffen A, Kales A, eds. *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. Washington, DC: National Institute of Health; 1968.

31. Dean DA, Goldberger AL, Mueller R, et al. Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource. *Sleep*. 2016;39(5):1151-1164. doi:10.5665/sleep.5774

32. Zhang G-Q, Cui L, Mueller R, et al. The National Sleep Research Resource: towards a sleep data commons. *J Am Med Informatics Assoc*. 2018;0(June):1-8. doi:10.1093/jamia/ocy064

33. Redline S, Sanders MH, Lind BK, et al. Methods for obtaining and analyzing unattended polysomnography data for a multicenter study. Sleep Heart Health Research Group. *Sleep*. 1998;21(7):759-767. http://www.ncbi.nlm.nih.gov/pubmed/11300121.

34. Quan SF, Howard B V, Iber C, et al. The Sleep Heart Health Study: design, rationale, and methods. *Sleep*. 1997;20(12):1077-1085. http://www.ncbi.nlm.nih.gov/pubmed/9493915.

35. Young T, Finn L, Peppard PE, et al. Sleep Disordered Breathing and Mortality: Eighteen-Year Follow-up of the Wisconsin Sleep Cohort. *Sleep*. 2008;31(8):291-292. doi:10.5665/sleep/31.8.1071

36. Young T, Palta M, Dempsey J, Skatrud J, Weber S, Badr S. The Occurrence of Sleep-Disordered Breathing among Middle-Aged Adults. *N Engl J Med*. 1993;328(17):1230-1235. doi:10.1056/NEJM199304293281704

37. Andlauer O, Moore H, Jouhier L, et al. Nocturnal Rapid Eye Movement Sleep Latency for Identifying Patients With Narcolepsy/Hypocretin

Deficiency. *JAMA Neurol*. 2013;70(7):891. doi:10.1001/jamaneurol.2013.1589

38. Moore H, Leary E, Lee S-Y, et al. Design and Validation of a Periodic Leg Movement Detector. *PLoS One*. 2014;9(12):e114565. doi:10.1371/journal.pone.0114565

39. Chambon S, Thorey V, Arnal PJ, Mignot E, Gramfort A, Neurospin CEA. A deep learning architecture to detect events in EEG signals during sleep. In: *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE; 2018:1-6.

40. Chambon S, Thorey V, Arnal PJ, Mignot E, Gramfort A. DOSED: A deep learning approach to detect multiple sleep micro-events in EEG signal. *J Neurosci Methods*. 2019;321:64-78. doi:10.1016/j.jneumeth.2019.03.017

41. Olesen AN, Chambon S, Thorey V, Jennum P, Mignot E, Sorensen HBD. Towards a Flexible Deep Learning Method for Automatic Detection of Clinically Relevant Multi-Modal Events in the Polysomnogram. In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE; 2019:556-561. doi:10.1109/EMBC.2019.8856570

42. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: *Proceedings of the 32nd International Conference on Machine Learning*. Vol 37. Lille, France: JMLR: W&CP; 2015. doi:10.1007/s13398-014-0173-7.2

43. He K, Zhang X, Ren S, Sun J. Identity Mappings in Deep Residual Networks. In: *Computer Vision -- ECCV 2016*. Vol abs/1603.0. ; 2016:630-645. doi:10.1007/978-3-319-46493-0_38

44. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ; 2015:770-778. doi:10.1109/CVPR.2016.90

45. Cho K, van Merrienboer B, Bahdanau D, Bengio Y. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2014:103-111. doi:10.3115/v1/W14-4012

46. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: *3rd International Conference on Learning Representations (ICLR)*. San Diego, CA; 2015:1-15. http://arxiv.org/abs/1412.6980.

47. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. *Proc Thirteen Int Conf Artif Intell Stat PMLR*. 2010;9:249-256.

48. Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch. In: *31st Conference on Neural Information Processing Systems (NIPS)*. Long Beach, CA, USA; 2017.

49. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.

50. Brink-Kjaer A, Olesen AN, Peppard PE, et al. *Automatic Detection of Cortical Arousals in Sleep and Their Contribution to Daytime Sleepiness*.; 2019. http://arxiv.org/abs/1906.01700.

51. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics*. 1977;33:159-174.

52. Sors A, Bonnet S, Mirek S, Vercueil L, Payen JF. A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomed Signal Process Control*. 2018;42:107-114. doi:10.1016/j.bspc.2017.12.001

53. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis*. 2015:211-252. doi:10.1007/s11263-015-0816-y

FIGURE CAPTIONS LIST

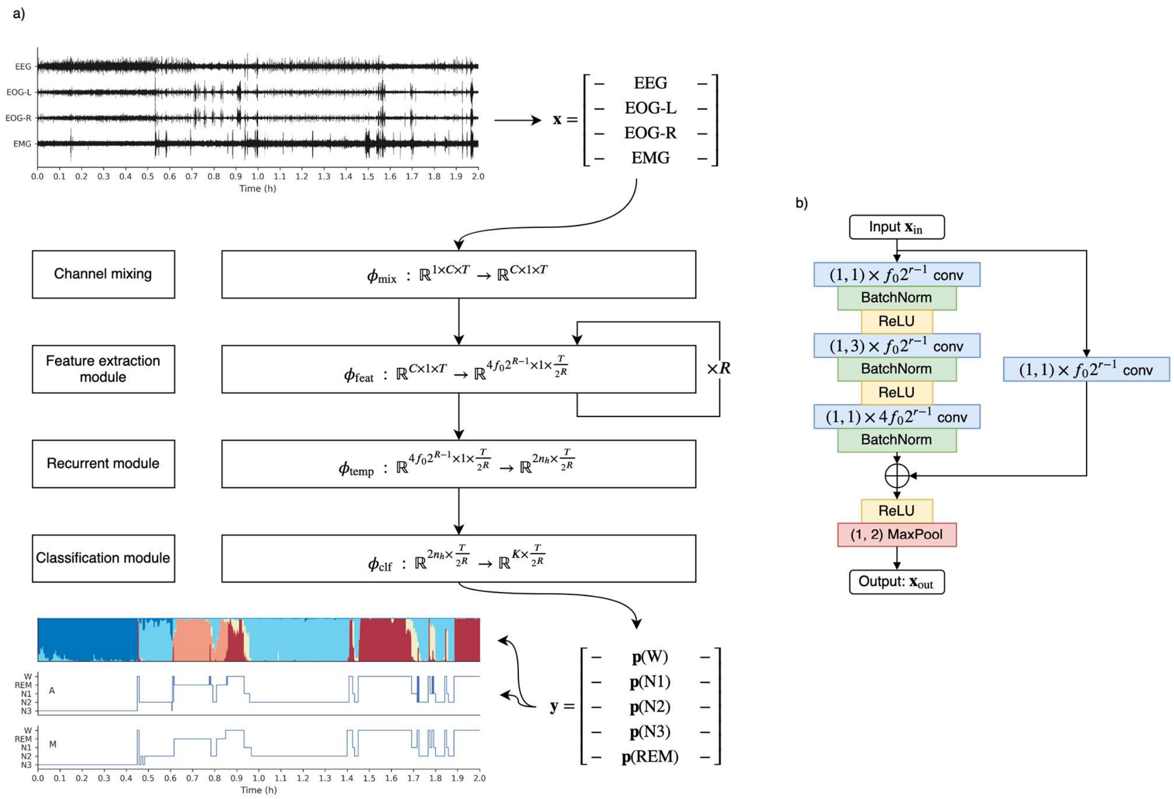

a)
EEG
EOG-L
EOG-R
EMG
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
Time (h)
$\mathbf{x} = \begin{bmatrix} - & \text{EEG} & - \\ - & \text{EOG-L} & - \\ - & \text{EOG-R} & - \\ - & \text{EMG} & - \end{bmatrix}$
Channel mixing
$\phi_{\text{mix}} : \mathbb{R}^{1\times C\times T} \rightarrow \mathbb{R}^{C\times 1\times T}$
Feature extraction module
$\phi_{\text{feat}} : \mathbb{R}^{C\times 1\times T} \rightarrow \mathbb{R}^{4f_0 2^{R-1}\times 1\times \frac{T}{2^R}}$
$\times R$
Recurrent module
$\phi_{\text{temp}} : \mathbb{R}^{4f_0 2^{R-1}\times 1\times \frac{T}{2^R}} \rightarrow \mathbb{R}^{2n_h\times \frac{T}{2^R}}$
Classification module
$\phi_{\text{clf}} : \mathbb{R}^{2n_h\times \frac{T}{2^R}} \rightarrow \mathbb{R}^{K\times \frac{T}{2^R}}$
W
REM
N1
N2
N3
A
M
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
Time (h)
$\mathbf{y} = \begin{bmatrix} - & \mathbf{p}(\text{W}) & - \\ - & \mathbf{p}(\text{N1}) & - \\ - & \mathbf{p}(\text{N2}) & - \\ - & \mathbf{p}(\text{N3}) & - \\ - & \mathbf{p}(\text{REM}) & - \end{bmatrix}$
b)
Input $\mathbf{x}_{\text{in}}$
$(1,1)\times f_0 2^{r-1}$ conv
BatchNorm
ReLU
$(1,3)\times f_0 2^{r-1}$ conv
BatchNorm
ReLU
$(1,1)\times 4f_0 2^{r-1}$ conv
BatchNorm
$(1,1)\times f_0 2^{r-1}$ conv
ReLU
(1, 2) MaxPool
Output: $\mathbf{x}_{\text{out}}$

Figure 1: Model overview. a) The input is a sequence of data $\mathbf{x}$ containing raw signal data from EEG, EOG-L/R, and EMG channels, which is supplied to the network modules in sequence. The feature extraction module consists of R repeated blocks of residual units, see panel to the right. The output of the model is a matrix $\mathbf{y}$ containing class probabilities for each sleep stage for each time step, which can be visualized either directly as a hypnodensity, or by $\arg\max \mathbf{y}$ as a hypnogram. The "A" and "M" labels in the hypnogram plots corresponds to automatic and manual hypnograms. b) Schematic of a single residual block in the feature extraction module. Convolutional layers are described by the kernel size × number of filters using a stride value of 1. Shortcut uses 1x1 convolutions with added zero-padding to maintain temporal dimension. Conv, convolutional layer; BatchNorm, batch normalization; ReLU, rectified linear unit; f_0, base number of filters ($f_0 = 4$).

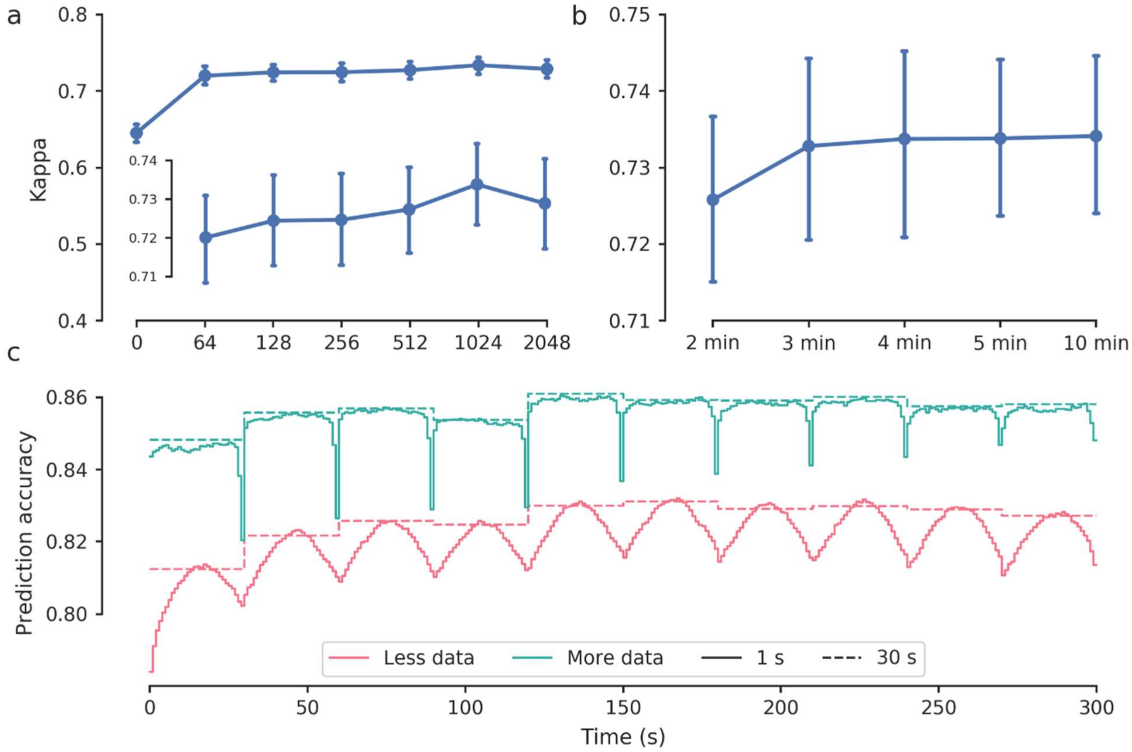


Figure 2: Temporal context changes model performance. a) Cohen's kappa as a function of the number of hidden units in the recurrent block. Inset shows zoom of Cohen's kappa for non-zero hidden unit values. b) Cohen's kappa as a function of sequence length. c) Prediction accuracy averaged across all 5-minute sequences in the test partition with a small and large training partition. Full lines are predictions evaluated every 1 s, while dashed lines show predictions averaged every 30 s. Values are shown for panels a), b) as mean with 95% confidence intervals.

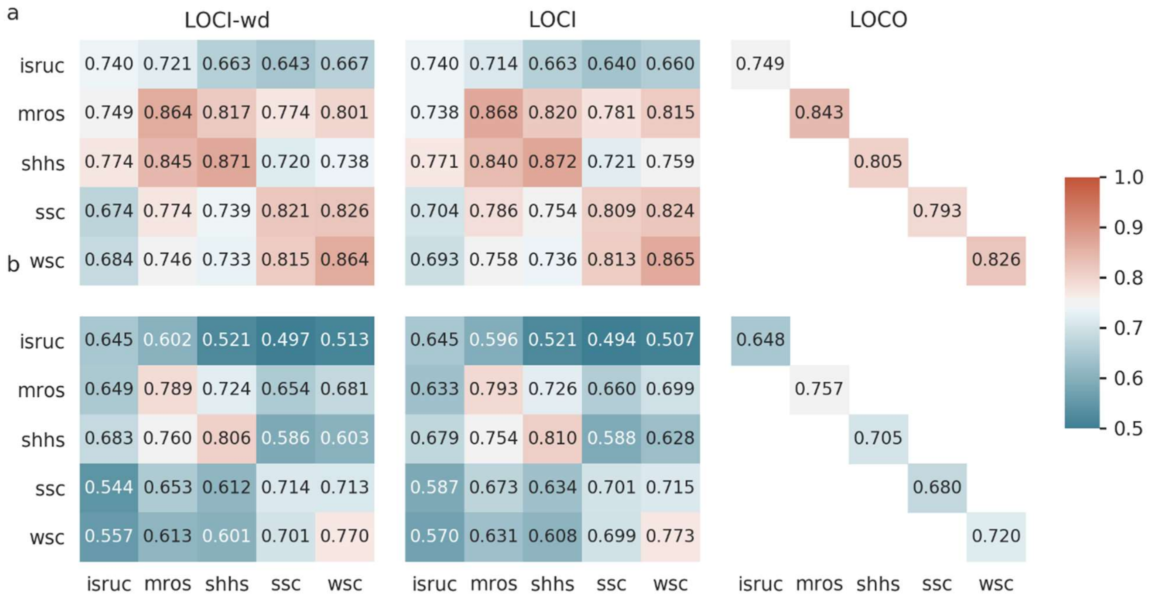


Figure 3: Individual cohorts influence classification performance on test partition ($N = 1{,}584$). As an example, training on MrOS in a LOCI configuration, the performance on the test subset of WSC is 0.815. The diagonals in all three configurations shows the performance for the same subjects in the test subsets in the respective cohorts making possible direct comparisons between LOCI and LOCO. For aggregated metrics and more summary statistics, please see ***Error! Reference source not found.***. LOCI, leave-one-cohort-in; LOCI-wd, LOCI with weight decay; LOCO, leave-one-cohort-out; ISRUC, Institute of Systems and Robotics, University of Coimbra Sleep Cohort; MrOS, MrOS Sleep Study; SHHS, Sleep Heart Health Study; SSC, Stanford Sleep Cohort; WSC, Wisconsin Sleep Cohort.

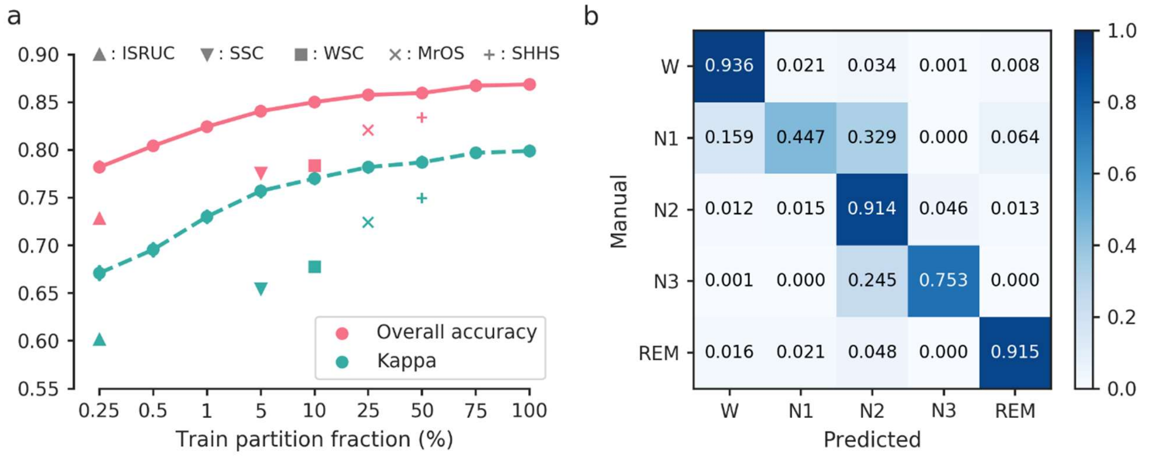


Figure 4: Training on mixed data increased predictive performance compared to individual cohorts of similar size. a) There is a gain in predictive performance by mixing data from various sources consistent across the size of the training dataset. b) Confusion matrix for a model trained on 100% of the available training partition data. The model shows excellent performance overall, with most misclassification happening between W and N1, and N1, N2, and N3. This is somewhat consistent with clinical experience, since N1 is a transition stage between wake and the deeper stages of sleep with much frequency content overlap with both W and N2.

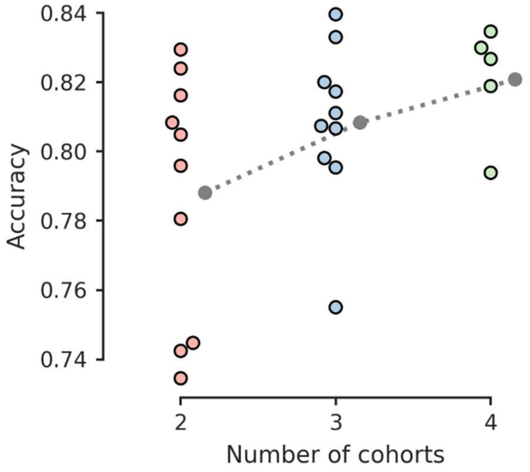


Figure 5: Number of cohorts in training partition increases model performance. Each datapoint is shown as the overall accuracy aggregated across all subjects for a specific training configuration. For example, the bottom dot in column 2 (3 cohort configuration) shows the performance on the test set (overall accuracy 0.755 ± 0.109, 95% CI: $[0.750 - 0.760]$), when training with 500 PSGs randomly and evenly drawn from the Stanford Sleep Cohort, the Institute of Systems and Robotics, University of Coimbra Sleep Cohort, and the Wisconsin Sleep Cohort. Notice the scale on the y-axis.

TABLES

Table 1: Cohort demographics.

	ISRUC	MrOS	SHHS	SSC	WSC	*p*-value
N (female)	126 (50)	3932 (0)	8444 (4458)	767 (319)	2401 (1103)	0
Age, years	49.8 ± 15.9 [20.0-85.0]	77.6 ± 5.6 [67.0-90.0]	64.5 ± 11.2 [39.0-90.0]	45.7 ± 14.5 [13.0-104.8]	59.7 ± 8.4 [37.2-82.3]	0
BMI, kg/m2	-	27.1 ± 3.8 [16.0-47.0]	28.2 ± 5.1 [18.0-50.0]	27.2 ± 6.5 [9.8-78.7]	31.6 ± 7.2 [17.5-70.6]	1.03e-171
TST, min	350.0 ± 67.3 [87.5-479.0]	352.1 ± 71.9 [39.0-626.0]	374.1 ± 69.4 [68.0-605.0]	361.0 ± 83.5 [0.0-661.0]	364.1 ± 63.6 [19.5-575.0]	4.07e-38
SL, min	17.7 ± 20.5 [0.0-144.5]	24.7 ± 26.9 [1.0-402.0]	24.2 ± 25.7 [0.0-349.0]	93.5 ± 58.9 [0.5-404.0]	33.2 ± 21.4 [0.5-333.0]	0
REML, min	125.6 ± 61.4 [7.0-323.0]	104.8 ± 75.1 [0.0-590.0]	91.7 ± 58.8 [0.0-471.0]	140.9 ± 88.0 [0.0-464.0]	128.3 ± 76.0 [3.5-514.0]	2.81e-173
WASO, min	76.2 ± 49.8 [7.5-251.0]	117.5 ± 67.6 [4.0-487.0]	80.2 ± 54.7 [2.0-378.0]	79.5 ± 55.0 [3.5-367.0]	73.6 ± 45.9 [3.0-325.0]	4.74e-233
SE, %	78.8 ± 14.1 [19.5-98.3]	75.5 ± 12.4 [12.0-99.0]	80.5 ± 11.0 [11.3-99.0]	77.4 ± 14.8 [0.0-98.0]	77.1 ± 11.2 [4.1-95.6]	4.23e-117
N1, %	13.3 ± 5.8 [1.8-33.1]	8.3 ± 6.4 [0.0-70.0]	5.5 ± 4.0 [0.0-39.1]	11.7 ± 10.2 [0.0-92.0]	10.8 ± 6.9 [1.0-88.4]	0
N2, %	31.9 ± 10.3 [4.4-89.3]	62.5 ± 10.0 [21.0-95.0]	56.9 ± 11.5 [10.9-100.0]	62.8 ± 24.9 [0.0-636.0]	66.0 ± 9.4 [9.1-93.3]	0
N3, %	19.6 ± 8.0 [0.0-41.1]	36.0 ± 31.8 [0.0-259.0]	17.5 ± 11.6 [0.0-70.1]	9.0 ± 9.3 [0.0-73.0]	7.2 ± 7.8 [0.0-47.5]	0
REM, %	13.3 ± 6.3 [0.0-37.8]	19.3 ± 6.8 [0.0-44.0]	20.1 ± 6.3 [0.0-48.0]	16.3 ± 7.2 [0.0-40.0]	16.0 ± 6.2 [0.0-38.2]	1.12e-203
ArI, /h	20.2 ± 10.0 [2.1-72.0]	23.7 ± 12.1 [1.0-105.0]	18.9 ± 10.5 [0.0-110.4]	125.0 ± 124.2 [1.0-729.0]	-	0
AHI, /h	13.1 ± 13.2 [0.0-82.2]	13.7 ± 14.6 [0.0-89.0]	18.1 ± 16.2 [0.0-161.8]	13.5 ± 19.2 [0.0-98.6]	7.0 ± 9.4 [0.0-72.6]	0
PLMI, /h	8.0 ± 27.4 [0.0-292.8]	35.7 ± 37.5 [0.0-233.0]	-	7.0 ± 18.1 [0.0-139.9]	-	1.22e-169

Cohort data represented as mean ± SD [range] unless noted. Arousal annotations were not available for WSC; PLMI was not available for SHHS and WSC; BMI was not available for ISRUC. N: number of subjects; TST: total sleep time; SL: sleep latency; REML: REM latency; WASO: wake after sleep onset; SE: sleep efficiency; ArI: arousal index; AHI: apnea/hypopnea index; PLMI: periodic leg movement index; ; ISRUC: Institute of Systems and Robotics, University of Coimbra Sleep Cohort; MrOS: The Osteoporotic Fractures in Men Sleep Study; SHHS: Sleep Heart Health Study; SSC: Stanford Sleep Cohort; WSC: Wisconsin Sleep Cohort.

Table 2: Overview of model architecture.

Module	**Type**	**# filters/units**	**Kernel size**	**Stride**	**Activation**	**Output size**
$\mathbf{x}$	Input	$-$	$-$	$-$	$-$	$1\ \times C\ \times T$
$\boldsymbol{\varphi}_{\text{mix}}$	2D convolution	C	$(1, C)$	1	$-$	$C\ \times 1\ \times T$
	Batch normalization	$-$	$-$	1	ReLU	$C\ \times 1\ \times T$
$\boldsymbol{\varphi}_{\text{feat}}^{(r)}$, $\boldsymbol{r} \in [\![\boldsymbol{R}]\!]$	†Residual module	$f_0 2^{r-1}/f_0 2^{r-1}/4f_0 2^{r-1}$	(1,1)/(1,3)/(1,1)	(1,1)/(1,2)	ReLU	$f_0 2^r \times 1 \times T/2^r$
$\boldsymbol{\varphi}_{\text{temp}}$	Bidirectional GRU	n_h	$-$	$-$	$-$	$2n_h \times T/2^R$
$\boldsymbol{\varphi}_{\text{clf}}$	1D convolution	K	$2n_h$	1	Softmax	$K\ \times T/2^R$

Kernel sizes correspond to the first, second and third convolutional layer in each residual block. Stride counts correspond to the residual block and the subsequent maximum pooling operation. ReLU, rectified linear unit; GRU, gated recurrent unit; C, number of input channels; T, length of segment in samples; f_0, base number of filters in residual blocks; R, number of residual blocks; n_h, number of hidden units in GRU; K, number of sleep stage classes. †SeeFigure 1.

SUPPLEMENTARY TABLES

Table S1: Temporal context impact on model performance in validation partition ($n = 426$).

	Overall accuracy				**Cohen's kappa**			
	Mean	**SD**	**Median**	**95% CI, mean**	**Mean**	**SD**	**Median**	**95% CI, mean**
Hidden units								
0	0.779	0.083	0.794	[0.771-0.787]	0.645	0.126	0.660	[0.633-0.657]
64	0.818	0.079	0.837	[0.810-0.825]	0.720	0.120	0.745	[0.709-0.731]
128	0.821	0.080	0.841	[0.813-0.829]	0.724	0.121	0.745	[0.713-0.736]
256	0.820	0.082	0.843	[0.812-0.828]	0.725	0.124	0.751	[0.713-0.736]
512	0.822	0.079	0.841	[0.815-0.830]	0.727	0.119	0.752	[0.716-0.739]
1024	0.828	0.072	0.845	[0.821-0.835]	0.734	0.111	0.758	[0.723-0.744]
2048	0.823	0.080	0.843	[0.816-0.831]	0.729	0.122	0.757	[0.717-0.740]
Sequence length								
2 min	0.821	0.075	0.840	[0.814-0.828]	0.726	0.114	0.754	[0.715-0.737]
3 min	0.826	0.080	0.845	[0.818-0.833]	0.733	0.123	0.762	[0.721-0.744]
4 min	0.828	0.079	0.849	[0.820-0.835]	0.734	0.122	0.762	[0.722-0.745]
5 min	0.828	0.072	0.845	[0.821-0.835]	0.734	0.111	0.758	[0.723-0.744]
10 min	0.829	0.075	0.848	[0.822-0.836]	0.734	0.113	0.759	[0.723-0.745]
Window length								
1 s	0.824	0.074	0.843	[0.817-0.831]	0.728	0.113	0.752	[0.717-0.738]
3 s	0.824	0.074	0.845	[0.817-0.832]	0.728	0.113	0.752	[0.717-0.739]
5 s	0.825	0.074	0.843	[0.818-0.832]	0.728	0.113	0.752	[0.717-0.739]
10 s	0.825	0.074	0.844	[0.818-0.832]	0.729	0.113	0.753	[0.718-0.739]
15 s	0.826	0.074	0.845	[0.818-0.833]	0.729	0.113	0.755	[0.719-0.740]
30 s	0.829	0.075	0.848	[0.822-0.836]	0.734	0.113	0.759	[0.723-0.745]

The **Hidden units** variable corresponds to varying the complexity in the recurrent module by increasing the number of hidden units. **Sequence length** indicate the length of the sequence of 30 epochs, while **Window length** correspond to varying the evaluation frequency.

Table S2: Performance characteristics for LOCI and LOCO training configurations.

	***N* PSGs**	**Overall accuracy**				**Cohen's kappa**			
		Mean	**SD**	**Median**	**95% CI, mean**	**Mean**	**SD**	**Median**	**95% CI, mean**
LOCI-wd									
ISRUC	1584	0.679	0.123	0.701	[0.673-0.685]	0.542	0.169	0.574	[0.533-0.550]
MrOS	1584	0.821	0.077	0.835	[0.817-0.825]	0.727	0.114	0.745	[0.721-0.733]
SHHS	1584	0.834	0.088	0.858	[0.830-0.839]	0.750	0.132	0.786	[0.744-0.757]
SSC	1584	0.762	0.094	0.774	[0.757-0.767]	0.639	0.129	0.654	[0.633-0.646]
WSC	1584	0.758	0.105	0.773	[0.753-0.764]	0.633	0.145	0.653	[0.626-0.640]
LOCI									
ISRUC	1584	0.676	0.124	0.700	[0.670-0.682]	0.539	0.170	0.574	[0.531-0.547]
MrOS	1584	0.826	0.074	0.839	[0.822-0.829]	0.732	0.111	0.748	[0.726-0.737]
SHHS‡	1584	0.837	0.084	0.858	[0.833-0.841]	0.754	0.127	0.786	[0.748-0.761]
SSC	1584	0.773	0.088	0.785	[0.769-0.777]	0.657	0.125	0.671	[0.651-0.663]
WSC	1584	0.763	0.101	0.776	[0.758-0.768]	0.641	0.140	0.659	[0.635-0.648]
LOCO									
ISRUC†	52	0.749	0.081	0.764	[0.727-0.771]	0.648	0.119	0.682	[0.616-0.680]
	126	*0.757*	*0.071*	*0.766*	*[0.744-0.769]*	*0.661*	*0.101*	*0.682*	*[0.643-0.678]*
MrOS†	371	0.843	0.066	0.851	[0.836-0.849]	0.757	0.104	0.776	[0.746-0.767]
	3932	*0.841*	*0.069*	*0.854*	*[0.838-0.843]*	*0.752*	*0.107*	*0.775*	*[0.749-0.755]*
SHHS	846	0.805	0.076	0.815	[0.800-0.810]	0.705	0.109	0.722	[0.698-0.712]
	8444	*0.800*	*0.081*	*0.811*	*[0.798-0.801]*	*0.697*	*0.115*	*0.713*	*[0.694-0.699]*
SSC	76	0.793	0.086	0.809	[0.744-0.812]	0.680	0.120	0.700	[0.653-0.707]
	766	*0.798*	*0.086*	*0.815*	*[0.792-0.805]*	*0.690*	*0.123*	*0.711*	*[0.681-0.699]*
WSC†	239	0.826	0.065	0.835	[0.818-0.834]	0.720	0.096	0.736	[0.708-0.732]
	2411	*0.824*	*0.068*	*0.837*	*[0.821-0.827]*	*0.718*	*0.100*	*0.736*	*[0.714-0.722]*

Metrics are aggregated across all subjects for each cohort in test partition ($N = 1{,}584$ PSGs). Statistics in italics correspond to evaluating performance on entire cohort. PSG: polysomnography; LOCI-wd: leave-one-cohort-in with weight decay; LOCO: leave-one-cohort-out; ISRUC: Institute of Systems and Robotics, University of Coimbra Sleep Cohort; MrOS: The Osteoporotic Fractures in Men Sleep Study; SHHS: Sleep Heart Health Study; SSC: Stanford Sleep Cohort; WSC: Wisconsin Sleep Cohort; †: significantly better than corresponding LOCI; ‡: significantly better than corresponding LOCO.

Table S3: Model performance of test partition with varying fractions of training data.

	Overall accuracy				Cohen's kappa			
	Mean	**SD**	**Median**	**95% CI, mean**	**Mean**	**SD**	**Median**	**95% CI, mean**
Fraction (%)								
0.25	0.782	0.097	0.801	[0.777-0.787]	0.671	0.141	0.696	[0.664-0.678]
0.50	0.804	0.086	0.824	[0.800-0.808]	0.696	0.131	0.724	[0.689-0.702]
1	0.824	0.079	0.840	[0.820-0.828]	0.730	0.118	0.753	[0.724-0.736]
5	0.841	0.074	0.856	[0.837-0.844]	0.757	0.113	0.780	[0.751-0.763]
10	0.850	0.069	0.864	[0.847-0.853]	0.770	0.108	0.791	[0.765-0.775]
25	0.858	0.066	0.873	[0.854-0.861]	0.782	0.102	0.804	[0.777-0.787]
50	0.860	0.063	0.874	[0.856-0.863]	0.787	0.097	0.809	[0.782-0.792]
75	0.867	0.062	0.882	[0.864-0.870]	0.797	0.096	0.818	[0.792-0.802]
100	0.869	0.064	0.883	[0.865-0.872]	0.799	0.098	0.820	[0.794-0.804]

Increasing the available training data increased performance on the test partition ($N = 1{,}584$) shown here as aggregated metrics across all subjects. No statistical difference was found by comparing confidence intervals (CI) between models trained with 75% and 100% of available training data, which indicates a saturation in training.

Table S4: Model performance on test partition ($N = 1{,}584$) with varying number of cohorts in training partition.

Training cohorts	**Overall accuracy**				**Kappa**			
	Mean	**SD**	**Median**	**95% CI, mean**	**Mean**	**SD**	**Median**	**95% CI, mean**
2								
Overall	0.788	0.102	0.811	[0.787-0.790]	0.683	0.143	0.710	[0.681-0.685]
ISRUC-MrOS	0.781	0.102	0.804	[0.776-0.786]	0.675	0.143	0.703	[0.668-0.682]
ISRUC-SHHS	0.808	0.097	0.835	[0.804-0.813]	0.717	0.142	0.756	[0.710-0.724]
ISRUC-SSC	0.735	0.103	0.753	[0.729-0.740]	0.613	0.140	0.638	[0.606-0.620]
ISRUC-WSC	0.745	0.107	0.758	[0.740-0.750]	0.628	0.140	0.642	[0.621-0.635]
MrOS-SHHS	0.829	0.081	0.849	[0.825-0.833]	0.740	0.124	0.769	[0.734-0.746]
MrOS-SSC	0.796	0.090	0.816	[0.791-0.800]	0.683	0.133	0.708	[0.677-0.690]
MrOS-WSC	0.805	0.087	0.822	[0.801-0.809]	0.699	0.126	0.722	[0.693-0.705]
SHHS-SSC	0.816	0.090	0.839	[0.812-0.821]	0.722	0.129	0.755	[0.716-0.729]
SHHS-WSC	0.824	0.089	0.846	[0.820-0.828]	0.733	0.128	0.762	[0.727-0.739]
SSC-WSC	0.742	0.110	0.755	[0.737-0.748]	0.620	0.145	0.634	[0.613-0.627]
3								
Overall	0.808	0.092	0.830	[0.807-0.810]	0.711	0.131	0.739	[0.709-0.713]
ISRUC-MrOS-SHHS	0.820	0.092	0.844	[0.815-0.825]	0.732	0.134	0.766	[0.725-0.738]
ISRUC-MrOS-SSC	0.798	0.088	0.816	[0.794-0.802]	0.694	0.129	0.720	[0.688-0.700]
ISRUC-MrOS-WSC	0.811	0.083	0.828	[0.807-0.815]	0.711	0.119	0.735	[0.705-0.717]
ISRUC-SHHS-SSC	0.807	0.090	0.828	[0.803-0.812]	0.714	0.126	0.739	[0.708-0.721]
ISRUC-SHHS-WSC	0.817	0.091	0.842	[0.813-0.822]	0.728	0.128	0.759	[0.722-0.735]
ISRUC-SSC-WSC	0.755	0.109	0.775	[0.750-0.760]	0.639	0.150	0.670	[0.631-0.646]
MrOS-SHHS-SSC	0.833	0.071	0.848	[0.829-0.837]	0.744	0.109	0.766	[0.739-0.750]
MrOS-SHHS-WSC	0.840	0.073	0.854	[0.836-0.843]	0.753	0.109	0.774	[0.748-0.759]
MrOS-SSC-WSC	0.795	0.088	0.811	[0.791-0.800]	0.687	0.123	0.706	[0.681-0.693]
SHHS-SSC-WSC	0.807	0.101	0.833	[0.802-0.812]	0.710	0.142	0.744	[0.703-0.717]
4								
Overall	0.821	0.085	0.840	[0.819-0.823]	0.728	0.124	0.755	[0.726-0.731]
ISRUC-MrOS-SHHS-SSC	0.827	0.078	0.843	[0.823-0.831]	0.739	0.115	0.764	[0.733-0.744]
ISRUC-MrOS-SHHS-WSC	0.835	0.075	0.850	[0.831-0.838]	0.747	0.112	0.768	[0.742-0.753]
ISRUC-MrOS-SSC-WSC	0.794	0.097	0.817	[0.789-0.799]	0.687	0.139	0.716	[0.680-0.694]
ISRUC-SHHS-SSC-WSC	0.819	0.091	0.843	[0.814-0.823]	0.728	0.131	0.759	[0.721-0.734]
MrOS-SHHS-SSC-WSC	0.830	0.076	0.846	[0.826-0.834]	0.741	0.112	0.763	[0.736-0.747]

The total number of training records were fixed at $N = 500$ for all configurations. ISRUC: Institute of Systems and Robotics, University of Coimbra Sleep Cohort; MrOS: The Osteoporotic Fractures in Men Sleep Study; SHHS: Sleep Heart Health Study; SSC: Stanford Sleep Cohort; WSC: Wisconsin Sleep Cohort.

C

PAPER III

TITLE: Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy

AUTHORS: Jens Stephansen, Alexander Neergaard Olesen, Mads Olsen, Aditya Ambati, Eileen B. Leary, Hyatt E. Moore IV, Oliver Carrillo, Ling Lin, Fang Han, Han Yan, Y L Sun, Yves Dauvilliers, Susan Scholz, L Barateau, Birgit Hogl, Ambra Stefani, S C Hong, T W Kim, Fabio Pizza, Giuseppe Plazzi, S Vandi, E Antelmi, Dimitri Perrin, Samuel T. Kuna, P. K. Schweitzer, Clete Kushida, Paul E. Peppard, Helge B. D. Sorensen, Poul Jennum, and Emmanuel Mignot

JOURNAL: Nature Communications

STATUS: Published

DOI: 10.1038/s41467-018-07229-3

FULL CITATION: Stephansen, J.B., Olesen, A.N., Olsen, M. *et al*. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nat Commun* **9**, 5229 (2018). https://doi.org/10.1038/s41467-018-07229-3

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-07229-3 OPEN

Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy

Jens B. Stephansen[1,2], Alexander N. Olesen[1,2,3], Mads Olsen[1,2,3], Aditya Ambati[1], Eileen B. Leary[1], Hyatt E. Moore[1], Oscar Carrillo[1], Ling Lin[1], Fang Han[4], Han Yan[4], Yun L. Sun[4], Yves Dauvilliers[5,6], Sabine Scholz[5,6], Lucie Barateau[5,6], Birgit Hogl[7], Ambra Stefani[7], Seung Chul Hong[8], Tae Won Kim[8], Fabio Pizza[9,10], Giuseppe Plazzi[9,10], Stefano Vandi[9,10], Elena Antelmi[9,10], Dimitri Perrin[11], Samuel T. Kuna[12], Paula K. Schweitzer[13], Clete Kushida[1], Paul E. Peppard[14], Helge B.D. Sorensen[2], Poul Jennum[3] & Emmanuel Mignot[1]

Analysis of sleep for the diagnosis of sleep disorders such as Type-1 Narcolepsy (T1N) currently requires visual inspection of polysomnography records by trained scoring technicians. Here, we used neural networks in approximately 3,000 normal and abnormal sleep recordings to automate sleep stage scoring, producing a hypnodensity graph—a probability distribution conveying more information than classical hypnograms. Accuracy of sleep stage scoring was validated in 70 subjects assessed by six scorers. The best model performed better than any individual scorer (87% versus consensus). It also reliably scores sleep down to 5 s instead of 30 s scoring epochs. A T1N marker based on unusual sleep stage overlaps achieved a specificity of 96% and a sensitivity of 91%, validated in independent datasets. Addition of HLA-DQB1*06:02 typing increased specificity to 99%. Our method can reduce time spent in sleep clinics and automates T1N diagnosis. It also opens the possibility of diagnosing T1N using home sleep studies.

[1] Center for Sleep Science and Medicine, Stanford University, Stanford 94304 CA, USA. [2] Department of Electrical Engineering, Technical University of Denmark, Kongens Lyngby 2800, Denmark. [3] Danish Center for Sleep Medicine, Rigshospitalet, Glostrup 2600, Denmark. [4] Department of Pulmonary Medicine, Peking University People's Hospital, Beijing 100044, China. [5] Sleep-Wake Disorders Center, Department of Neurology, Gui-de-Chauliac Hospital, CHU Montpellier 34295, France. [6] INSERM, U1061, Université Montpellier 1, Montpellier 34090, France. [7] Department of Neurology, Innsbruck Medical University, Innsbruck 6020, Austria. [8] Department of Psychiatry, St. Vincent's Hospital, The Catholic University of Korea, Seoul 16247, Korea. [9] Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna 40123, Italy. [10] IRCCS Istituto delle Scienze Neurologiche di Bologna, Bologna 40139, Italy. [11] School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane 4001, Australia. [12] Department of Medicine and Center for Sleep and Circadian Neurobiology, University of Pennsylvania, Philadelphia 19104 PA, USA. [13] Sleep Medicine and Research Center, St. Luke's Hospital, Chesterfield 63017 MO, USA. [14] Department of Population Health Sciences, University of Wisconsin-Madison, Madison 53726 WI, USA. These authors contributed equally: Jens B. Stephansen, Alexander N. Olesen. These authors jointly supervised this work: Helge B.D. Sorensen, Poul Jennum, Emmanuel Mignot. Correspondence and requests for materials should be addressed to E.M. (email: mignot@stanford.edu)

Sleep disorders and sleep dysregulation impact over 100 million Americans, contributing to medical consequences such as cardiovascular (arrhythmia, hypertension, stroke), metabolic (diabetes, obesity) and psychiatric disorders (depression, irritability, addictive behaviors). Sleep deprivation impairs performance, judgment and mood, and is a major preventable contributor to accidents[1]. There are ~90 different sleep disorders including insomnia (20% of population), obstructive and central sleep apnea (10%), restless legs syndrome (4%), rapid eye movement (REM) sleep behavior disorder (RBD) and hypersomnia syndromes such as type 1 narcolepsy (T1N)[2].

Among these pathologies, T1N is unique as a disorder with a known, discrete pathophysiology—a destruction of hypocretin neurons in the hypothalamus likely of autoimmune origin[3]. This is reflected in the cerebrospinal fluid (CSF) concentrations of the hypocretin-1 (orexin-A) neuropeptide, where a concentration below 110 pg/ml is considered indicative of narcolepsy[2]. Typically beginning in childhood or adolescence, narcolepsy affects approximately 0.03% of the US, European, Korean and Chinese populations[4]. Unique to narcolepsy is the extremely strong (97% versus 25%) association with a genetic marker, HLA-DQB1*06:02[5], and a well-characterized set of sleep disturbances that include short sleep latency, rapid transitions into REM sleep and poor nocturnal sleep consolidation. The pathology also includes episodes of "sleep/wake dissociation" where the patient is half awake and half in REM sleep, for example, experiencing REM sleep muscle paralysis while awake (sleep paralysis, cataplexy) or dreaming while awake (hypnagogic hallucinations).

Sleep disorders are generally assessed at sleep clinics by performing sleep analysis using nocturnal polysomnography (PSG), a recording comprised of multiple digital signals which include electroencephalography (EEG), electrooculography (EOG), chin and leg electromyography (EMG), electrocardiography, breathing effort, oxygen saturation and airflow[6]. When sleep is analyzed in PSGs, it is divided into discrete stages: wake, non-REM (NREM) sleep stage 1 (N1), 2 (N2) and 3 (N3), and REM. Each stage is characterized by different criteria, as defined by consensus rules published in the American Academy of Sleep Medicine (AASM) Scoring Manual[6,7]. N1 (sleep onset) is characterized by slowing of the EEG, disappearance of occipital alpha waves, decreased EMG and slow rolling eye movements, while N2 is associated with spindles and K-complexes. N3 is characterized by a dominance of slow, high amplitude waves (>20%), while REM sleep is associated with low voltage, desynchronized EEG with occasional saw tooth waves, low muscle tone and REMs. PSG analysis is typically done by certified technicians who, through visual inspection on a standardized screen, assign a sleep stage to each 30 s segment of the full recording. Although there is progression from N1 to N3 then to REM during the night, a process that repeats approximately every 90 min (the sleep cycle), each stage is associated with physiological changes that can be meaningful to the assessment of sleep disorders such as obstructive sleep apnea. For example, the abnormal breathing events that occur with obstructive sleep apnea (OSA) are generally less severe in N3 versus N2 because of central control of breathing changes, and they are more severe in REM sleep, due to upper airway muscle weakness[8]. The differentiation of sleep stages is also particularly important for the diagnosis of narcolepsy, a condition currently assessed by a PSG followed by a multiple sleep latency test (MSLT), a test where patients are asked to nap 4–5 times for 20 min every 2 h during the daytime and sleep latency and the presence of REM sleep is noted[9]. A mean sleep latency (MSL) less than 8 min (indicative of sleepiness) and the presence of at least 2 sleep onset REM periods (SOREMPs, REM latency ≤15 min following sleep onset in naps) during the MSLT or 1 SOREM plus a REM latency ≤15 min during nocturnal PSG is diagnostic for narcolepsy. In a recent large study of the MSLT[10], specificity and sensitivity for type 1 narcoleptics were, respectively, 98.6% and 92.9% in comparing 516 T1N versus 516 controls and 71.2% and 93.4% in comparing 122 T1N cases versus 132 other hypersomnia cases (high pretest probability cohort). Similar sensitivity (75–90%) and specificity (90–98%) have been reported by others in large samples of hypersomnia cases versus T1N[11–15].

Manual inspection of sleep recordings has many problems. It is time consuming, expensive, inconsistent, subjective and must generally be done offline. In one study, Rosenberg and Van Hout[16] found inter-scorer reliability for sleep stage scoring to be 82.6% on average, a result consistently found by others[17–20]. N1 and N3 in particular have agreements as low as 63 and 67%, placing constraints on their usefulness[16]. In this study, we explored whether deep learning, a specific subtype of machine learning, could produce a fast, inexpensive, objective, and reproducible alternative to manual sleep stage scoring. In recent years, similar complex problems such as labeling images, understanding speech and translating language have seen advancement to the point of outperforming humans[21–23]. Several high-profile papers have also documented the efficacy of deep learning algorithms in the healthcare sector, especially in the fields of diabetic retinopathy[24,25], digital pathology[26,27] and radiology[28,29]. This technology refers to complex neural network models with a very large number (on a magnitude of millions) of parameters and processing layers. For a thorough review of the underlying theory behind deep learning including common model paradigms, we refer to the review article by LeCun et al.[30].

In this implementation of deep learning, we introduce the hypnodensity graph—a hypnogram that does not enforce a single sleep stage label, but rather a membership function to each of the sleep stages, allowing more information about sleep trends to be conveyed, something that is only possible in non-human scoring. Using this concept, we next applied deep learning-derived hypnodensity features to the diagnosis of T1N, showing that an analysis of a single PSG night can perform as well as the PSG-MSLT gold standard, a 24 h long procedure.

Results

Inter-scorer reliability cohort. Supplementary Table 1 reports on the description of the various cohorts included in this study, and how they were utilized (see Datasets section in Methods). These originate from seven different countries. We assessed inter-scorer reliability using the Inter-scorer Reliability Cohort (IS-RC)[31], a cohort of 70 PSGs scored by 6 scorers across three locations in the United States[31]. Table 1 displays individual scorer performance as well as the averaged performance across scorers, with top and bottom of table showing accuracies and Cohen's kappas, respectively. The results are shown for each individual scorer when compared to the consensus of all scorers (biased) and compared to the consensus of the remaining scorers (unbiased). In the event of no majority vote for an epoch, the epoch was counted equally in all classes in which there was disagreement. Also shown in Table 1 is the model performance on the same consensus scorings as each individual scorer along with the *t*-statistic and associated *p* value for each paired *t*-test between the model performance and individual scorer performance. At a significance level of 5%, the model performs statistically better than any individual scorer both in terms of accuracy and Cohen's kappa.

Supplementary Table 2 displays the confusion matrix for every epoch of every scorer of the inter-scorer reliability data, both unadjusted (top) and adjusted (bottom). As in Rosenberg and Van Hout[16], the biggest discrepancies occur between N1 and

Table 1 Individual and overall scorer performance, expressed as accuracy and Cohen's kappa

	Overall	Scorer 1	Scorer 2	Scorer 3	Scorer 4	Scorer 5	Scorer 6
Accuracy (%), biased	81.3 ± 3.0	82.4 ± 6.1	84.6 ± 5.5	74.1 ± 7.9	85.4 ± 5.7	83.1 ± 9.4	78.3 ± 8.9
Accuracy (%), unbiased	76.0 ± 3.2	77.3 ± 6.3	79.1 ± 6.3	69.0 ± 8.0	79.7 ± 6.5	77.8 ± 9.6	72.9 ± 9.2
Model accuracy (%) on concensus	—	85.1 ± 4.9	83.8 ± 5.0	86.5 ± 4.3	84.3 ± 4.7	85.6 ± 4.7	87.0 ± 4.5
T-stat (*p* value)	—	9.5 (3.8×10^{-14})	6.6 (7.5×10^{-9})	18.3 (6.0×10^{-28})	6.7 (4.7×10^{-9})	6.4 (1.7×10^{-8})	12.2 (7.5×10^{-19})
Cohen's kappa, biased	61.0 ± 6.8	63.6 ± 12.2	68.4 ± 10.5	45.6 ± 19.7	69.6 ± 13.2	64.5 ± 20.9	54.5 ± 19.8
Cohen's kappa, unbiased	57.7 ± 6.1	61.3 ± 11.2	64.6 ± 10.3	43.5 ± 19.2	64.6 ± 13.1	60.9 ± 16.9	51.6 ± 16.7
Model kappa on concensus	—	74.3 ± 12.3	72.4 ± 12.1	76.0 ± 11.8	72.7 ± 12.0	74.7 ± 12.1	76.6 ± 12.2
T-stat (*p* value)	—	9.5 (4.6×10^{-14})	7.1 (7.9×10^{-10})	15.4 (7.0×10^{-24})	6.6 (6.4×10^{-9})	7.1 (9.2×10^{-10})	13.2 (2.0×10^{-20})

Both accuracy and Cohen's kappa are presented as both with (biased) and without (unbiased) the assessed scorer included in the consensus standard in a leave-one-out fashion. Accuracy is expressed in percent, and Cohen's kappa is a ratio, and therefore unitless. *T*-statistics and *p* values correspond to the paired *t*-test between the unbiased predictions for each scorer against the model predictions on the same consensus

Wake, N1 and N2, and N2 and N3, with some errors also occurring between N1 and REM, and N2 and REM.

For future analyses of the IS-RC in combination with other cohorts that have been scored only by one scorer, a final hypnogram consensus was built for this cohort based on the majority vote weighted by the degree of consensus from each voter, expressed as its Cohen's κ, $\boldsymbol{\kappa} = 1 - \frac{1-p_o}{1-p_e}$, where p_e is the baseline accuracy and p_o is the scorer accuracy, such that

$$\mathbf{y} = \arg\max \frac{\sum_{i=1}^{6} \widehat{\mathbf{y}}_i \cdot \boldsymbol{\kappa}_i}{\sum_{i=6}^{6} \boldsymbol{\kappa}_i}. \tag{1}$$

In this implementation, scorers with a higher consensus with the group are considered more reliable and have their assessments weighted heavier than the rest. This also avoided split decisions on end-results.

Optimizing machine learning performance for sleep staging. We next explored how various machine learning algorithms (see Methods) performed depending on cohort, memory (i.e., feed forward (FF) versus long short-term memory networks (LSTM)), signal segment length (short segments of 5 s (SS) versus long segments of 15 s (LS)), complexity (i.e., low (SH) vs. high (LH)), encoding (i.e., octave versus cross-correlation (CC) encoding, and realization type (repeated training sessions). The performance of these machine learning algorithms was compared with the six-scorer consensus in the IS-RC and with single scorer data in 3 other cohorts, the Stanford Sleep Cohort (SSC)[10,32], the Wisconsin Sleep Cohort (WSC)[32,33] and the Korean Hypersomnia Cohort (KHC)[10,34] (see Datasets section in Methods for description of each cohort).

Model accuracy varies across datasets, reflecting the fact scorer performance may be different across sites, and because unusual subjects such as those with specific pathologies can be more difficult to score—a problem affecting both human and machine scoring. In this study, the worst performance was seen in the KHC and SSC with narcolepsy, and the best performance was achieved on IS-RC data (Supplementary Figure 1a, Table 2, Supplementary Table 7). The SSC+KHC cohorts mainly contain patients with more fragmented sleeping patterns, which would explain a reduced performance. The IS-RC has the most accurate label, minimizing the effects of erroneous scoring, which therefore leads to an increased performance. Incorporating large ensembles of different models increased mean performance slightly (Table 2).

The two most important factors that increased prediction accuracy were encoding and memory, while segment length, complexity and number of realizations were less important (Supplementary Figure 1). The effect of encoding was less prominent in the IS-RC. Prominent factor interactions include (Supplementary Figure 2): (i) CC encoding models improve with higher complexity, whereas octave encoding models worsen; (ii) increasing segment length positively affects models with low complexity, but does not affect models with a high complexity; and (iii) adding memory improves models with an octave encoding more than models with a CC encoding. Because the IS-RC data are considered the most reliable, we decided to use these data as benchmark for model comparison. This standard improved as more scorers were added, and the model performance increased. (Fig. 1a). The different model configurations described in this section do not represent exhaustive configuration search, and future work experiments might result in improved results.

Figure 2a displays typical scoring outputs (bottom panels) obtained with a single sleep study of the IS-RC cohort in comparison to 6 scorer consensus (top panel). The model results are displayed as hypnodensity graphs, representing not only discrete sleep stage outputs, but also the probability of occurrence of each sleep state for each epoch (see definition in Data labels, scoring and fuzzy logic section). As can be seen, all models performed well, and segments of the sleep study with the lowest scorer consensus (top) are paralleled by similar sleep stage probability uncertainty, with performance closest to scoring consensus achieved by an ensemble model described below (second to top).

Final implementation of automatic sleep scoring algorithm. Because of model noise, potential inaccuracies and the desire to quantify uncertainty, the final implementation of our sleep scoring algorithm is an ensemble of different CC models with small variations in model parameters, such as the number of feature-maps and hidden nodes. This was achieved by randomly varying the parameters between 50 and 150% of the original values using the CC/SH/LS/LSTM as a template (this model achieved similar performance to the CC/LH/LS/LSTM while requiring significantly less computational power).

All models make errors, but as these errors occur independently of each other, the risk of not detecting and correcting errors falls with increasing model numbers. For this reason, 16 such models were trained, and at each analyzed segment both mean and variance of model estimates were calculated. As expected, the relative model variance (standardized to the average variance in a correct wakefulness prediction) is generally lower in correct predictions (Supplementary Table 3) and this can be used to inform users about uncertain/incorrect estimates. To demonstrate the effectiveness of this final implementation, the average of the models is shown alongside the distribution of 5234 ± 14 scorers on 150 epochs, a dataset provided by the AASM (AASM inter-scorer reliability (ISR) dataset, (see Datasets section

Table 2 Performance of best models, as they are described by Supplementary Table 8, on various datasets compared to the six-scorer consensus

Test data	Best single model	Mean performance (%)	Best ensemble	Mean performance (%)
WSC	CC/SH/LS/LSTM/2	86.0 ± 5.0	All CC	86.4 ± 5.2
SSC+KHC, no narcolepsy	CC/LH/SS/LSTM	76.9 ± 11.1	All CC	77.0 ± 11.9
SSC+KHC, narcolepsy	CC/LH/SS/LSTM	68.8 ± 11.0	All CC	68.4 ± 12.2
IS-RC	CC/LH/LS/LSTM/2	84.6 ± 4.6	All models	86.8 ± 4.3

All comparisons are on a by-epoch basis

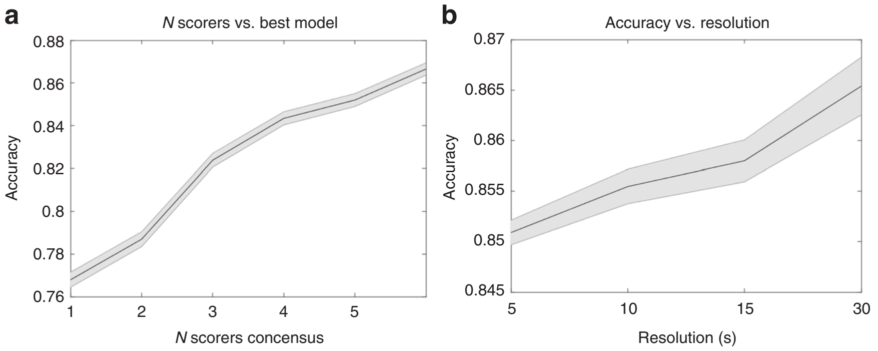


Fig. 1 Accuracy per scorer and by time resolution. **a** The effect on scoring accuracy as golden standard is improved. Every combination of *N* scorers is evaluated in an unweighted manner and the mean is calculated. Accuracy is shown with mean (solid black line) and a 95% confidence interval (gray area). **b** Predictive performance of best model at different resolutions. Performance is shown as mean accuracy (solid black line) with a 95% confidence interval (gray area)

in Methods). On these epochs, the AASM ISR achieved a 90% agreement between scorers. In comparison, the model estimates reached a 95% accuracy compared to the AASM consensus (Fig. 2b). Using the model ensemble and reporting on sleep stage probabilities and inter-model variance for quality purpose constitute the core of our sleep scoring algorithm.

Ensemble/best model performance. Supplementary Table 2 reports on concordance for our best model, the ensemble of all CC models. Concordance is presented in a weighted and unweighted manner, between the best model estimate and scorer consensus (Table 3). Weighing of a segment was based on scorer confidence and serves to weigh down controversial segments. For each recording i, the epoch-specific weight ω_n and weighted accuracy α_ω were calculated as:

$$\omega_n = \max_{z\in\mathcal{Z}}\left(\mathbf{P}\left(\mathbf{y}_n|\mathbf{x}_n\right)_z\right) - \ell^2_{\mathcal{Z}}\left(\mathbf{P}\left(\mathbf{y}_n|\mathbf{x}_n\right)\right), \tag{2}$$

$$\alpha^{(i)}_\omega = \frac{1}{\sum_n \omega_n}\sum_n \omega_n \cdot \left(\arg\max_{m\in\mathcal{M}}\left(\mathbf{P}_m\left(\hat{\mathbf{y}}_n|\mathbf{x}_n\right)\right) \cap \arg\max_{z\in\mathcal{Z}}\left(\mathbf{P}_z\left(\mathbf{y}_n|\mathbf{x}_n\right)\right)\right), \tag{3}$$

where $\ell^2_{\mathcal{Z}}\left(\mathbf{P}\left(\mathbf{y}_n|\mathbf{x}_n\right)\right)$ is the second most likely stage assessed by the set of scorers (experts) denoted by $\mathcal{Z}$, of the n^{th} epoch in a sleep recording. As with scorers, the biggest discrepancies occurred between wake versus N1, N1 versus N2 and N2 versus N3. Additionally, the weighted performance was almost universally better than the unweighted performance, raising overall accuracy from 87 to 94%, indicating a high consensus between automatic scoring and scorers in places with high scorer confidence. An explanation for these results could be that both scorers and model are forced to make a choice between two stages when data are ambiguous. An example of this may be seen in Fig. 2a. Between 1 and 3 h, several bouts of N3 occur, although they often do not reach the threshold for being the most likely stage. As time progresses, more evidence for N3 appears reflecting increased proportion of slow waves per epoch, and confidence increases, which finally yields “definitive” N3. This is seen in both model and scorer estimates. Choosing to present the data as hypnodensity graphs mitigates this problem. The various model estimates produce similar results, which also resemble the scorer assessment distribution, although models without memory fluctuate slightly more, and tend to place a higher probability on REM sleep in periods of wakefulness, since no contextual information is provided.

Influences of sleep pathologies. As seen in Table 2, the different cohorts achieve different performances. To see how much may be attributed to various pathologies, five different analyses of variance were made, with accuracy as the dependent variable, using cohort, age (grouped as age < 30, 30 ≤ age < 50 and age ≥ 50) and sex as covariates (Supplementary Table 4), investigating the effect of insomnia, OSA, restless leg syndrome (RLS), periodic leg movement index (PLMI) and T1N on accuracy of our machine learning routine versus human scoring. This was performed in the cohort mentioned above with addition of the Austrian Hypersomnia Cohort (AHC)[35]. The p values obtained from paired t-testing for each condition were 0.75 (insomnia), 7.53×10^{-4} (OSA), 0.13 (RLS), 0.22 (PLMI) and 1.77×10^{-15} (T1N) respectively, indicating that only narcolepsy had a strong effect on scorer performance. Additionally, in the context of narcolepsy, cohort and age yielded p values between 3.69×10^{-21} and 2.81×10^{-82} and between 0.62 and 6.73×10^{-6}, respectively. No significant effect of gender was ever noted. Cohort effects were expected and likely reflect local scorer performances and differences in PSG hardware and filter setups at every site. Decreased performance with age likely reflects decreased EEG amplitude, notably in N3/slow wave sleep amplitude with age[36].

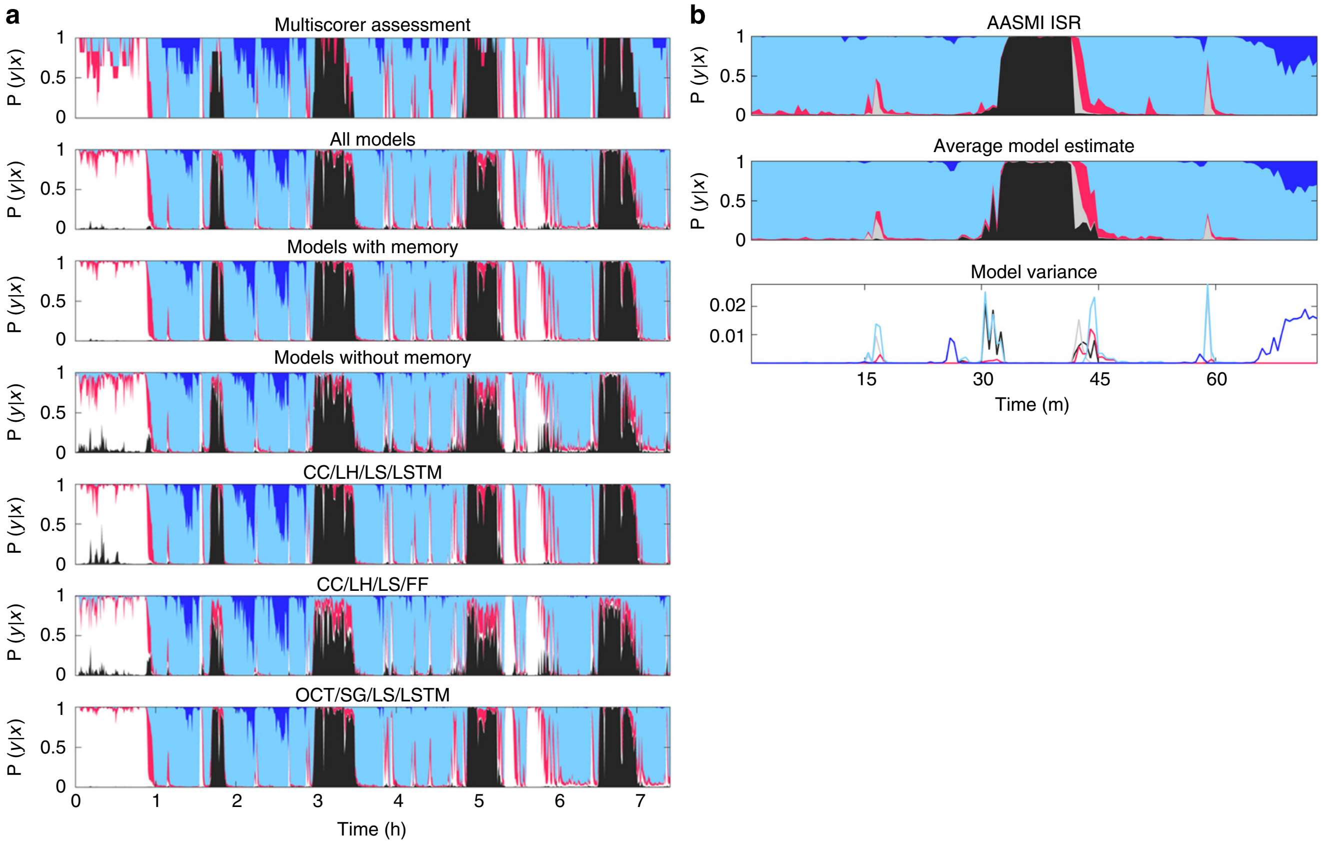


Fig. 2 Hypnodensity example evaluated by multiple scorers and different predictive models. **a** The figure displays the hypnodensity graph. Displayed models are, in order: multiple scorer assessment (1); ensembles as described in Supplementary Table 8: All models, those with memory (LSTM) and those without memory (FF) (2–4); single models, as described in Supplementary Table 8 (5–7). OCT is octave encoding, Color codes: white, wake; red, N1; light blue, N2; dark blue, N3; black, REM. **b** The 150 epochs of a recording from the AASM ISR program are analyzed by 16 models with randomly varying parameters, using the CC/SH/LS/LSTM model as a template. These data were also evaluated by 5234 ± 14 different scorers. The distribution of these is shown on top, the average model predictions are shown in the middle, and the model variance is shown at the bottom

Table 3 Confusion matrix displaying the relation between different targets and the ensemble estimate

	Target					
Model Predictions	**Wake**	**N1**	**N2**	**N3**	**REM**	**Precision**
Wake	14.08%	0.35%	0.88%	0.007%	0.08%	0.91
	16.68%	0.15%	0.44%	0.003%	0.02%	0.96
N1	1.13%	1.78%	3.00%	0.002%	0.36%	0.28
	0.47%	0.88%	1.15%	0%	0.12%	0.34
N2	0.29%	0.59%	52.58%	1.27%	0.66%	0.95
	0.12%	0.25%	56.30%	0.34%	0.32%	0.98
N3	0.002%	0%	2.13%	4.87%	0%	0.70
	0%	0%	1.09%	4.23%	0%	0.91
REM	0.54%	1.17%	0.78%	0%	13.45%	0.84
	0.40%	0.73%	0.41%	0%	15.86%	0.91
Sensitivity	0.88	0.46	0.89	0.79	0.92	0.87
	0.94	0.44	0.95	0.92	0.97	0.94

The targets are: top row: unweighted consensus; bottom row: weighted by the scorer agreement at each epoch. The number of analyzed epochs were 53,009 (unweighted) and 36,032 (weighted)

Resolution of sleep stage scoring. Epochs are evaluated with a resolution of 30 s, a historical standard that is not founded in anything physiological, and limits the analytical possibilities of a hypnogram. Consequently, it was examined to what extent the performance would change as a function of smaller resolution. Only the models using a segment size of 5 s were considered. Segments were averaged to achieve performances at 5, 10, 15 and 30 s resolutions, and the resulting performances in terms of accuracy are shown in Fig. 1b. Although the highest performance was found using a resolution of 30 s, performance dropped only slightly with decreasing window sizes.

Construction and evaluation of a narcolepsy biomarker. The neural networks produce outputs that depend on evidence in the input data for or against a certain sleep stage based on features learned through training. We hypothesized that narcolepsy, a condition characterized by sleep/wake stage mixing/dissociation[37–41], would result in a greater than normal overlap between stages, an observation that was obvious when sleep stage probability were plotted in such subjects (see example in Fig. 3). Based on this result, we hypothesized that such sleep stage model outputs could be used as a biomarker for the diagnosis of narcolepsy using a standard nocturnal PSG rather than the more time-consuming MSLT.

To quantify narcolepsy-like behavior for a single recording, we generated features quantifying sleep stage mixing/dissociation. These are based on descriptive statistics and other features describing persistence of a set of new time series generated from the geometric mean of every permutation of the set of sleep stages, as obtained from the 16 CC sleep stage prediction models.

In addition to this, we also added features expected to predict narcolepsy based on prior work, such as REM sleep latency and sleep stage sequencing parameters (see “Hypnodensity as feature for the diagnosis of T1N” section in Methods for details). A recursive feature elimination (RFE) procedure[42] was performed on extracted features with average outcome putting the optimal number of relevant features at 38. An optimal selection frequency cut-off of 0.40 (i.e., including a feature if it was selected 40% of the time) was determined using a cross-validation setup on the training data. Features are described in Supplementary Table 5 with detailed description of the 8 most important features reported in Table 4.

Final predictions were achieved by creating a separate Gaussian Predictor (GP) narcolepsy classifier from each of the sleep scoring models used in the final implementation. This was tested in seven independent datasets: a training dataset constituted of PSG from WSC[32,33], SSC[10,32], KHC[10,34], AHC[35], Jazz Clinical Trial Sample (JCTS)[43], Italian Hypersomnia Cohort (IHC)[41] and DHC; with verification in test data mostly constituted of PSG from the same cohorts and independent replication in the French Hypersomnia Cohort (FHC) and the Chinese Narcolepsy Cohort (CNC)[12] that had never been seen by the algorithm (see Supplementary Table 1). The algorithm produced values between −1 and 1, with 1 indicating a high probability of narcolepsy. A cut-off threshold between narcolepsy type 1 and “other“ was set at −0.03 (red dot, Fig. 4), determined using training data, as shown in Fig. 4a. The optimal trade-off achieves both high sensitivity and specificity, which is seen to translate well onto the test data (Fig. 4b) and the never seen replication sample (Fig. 4c).

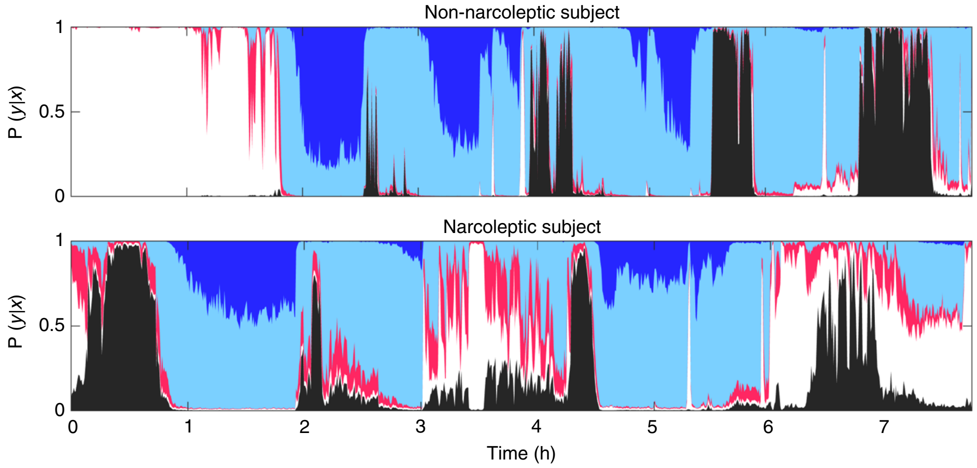


Fig. 3 Examples of hypnodensity graph in subjects with and without narcolepsy. Hypnodensity, i.e., probability distribution per stage of sleep for a subject without narcolepsy (top) and a subject with narcolepsy (Bottom). Color codes: white, wake; red, N1; light blue, N2; dark blue, N3; black, REM

Table 4 Descriptions of the 8 most frequently selected features

Number	Relative selection frequency	Description
1	1	The time taken before 5% of the sum of the product between W, N2 and REM, calculated at every epoch, has accumulated, weighed by the total amount of this sum. This feature expresses the known sleep stage dissociation and altered sleep timing.
2	0.91	The number of nightly SOREMPS appearing throughout the recording.
3	0.82	The time taken before 50% of the wakefulness in a recording has accumulated, weighed by the total amount of wakefulness.
4	0.82	REM 6 The Shannon entropy of the REM sleep stage distribution. This expresses the amount of information held in a signal, or in this case, how many different values the REM sleep stage distribution obtains—how consolidated phases of REM are when the stage appears.
5	0.68	The maximum probability of wakefulness obtained in a recording.
6	0.68	The maximum value obtained of the product between the N2 and REM probability in a recording.
7	0.68	The time taken before 30% of the sum of the product between W and N2, calculated at every epoch, has accumulated, weighed by the total amount of this sum.
8	0.64	The time taken before 10% of the sum of the product between W and N1, calculated at every epoch, has accumulated, weighed by the total amount of this sum.

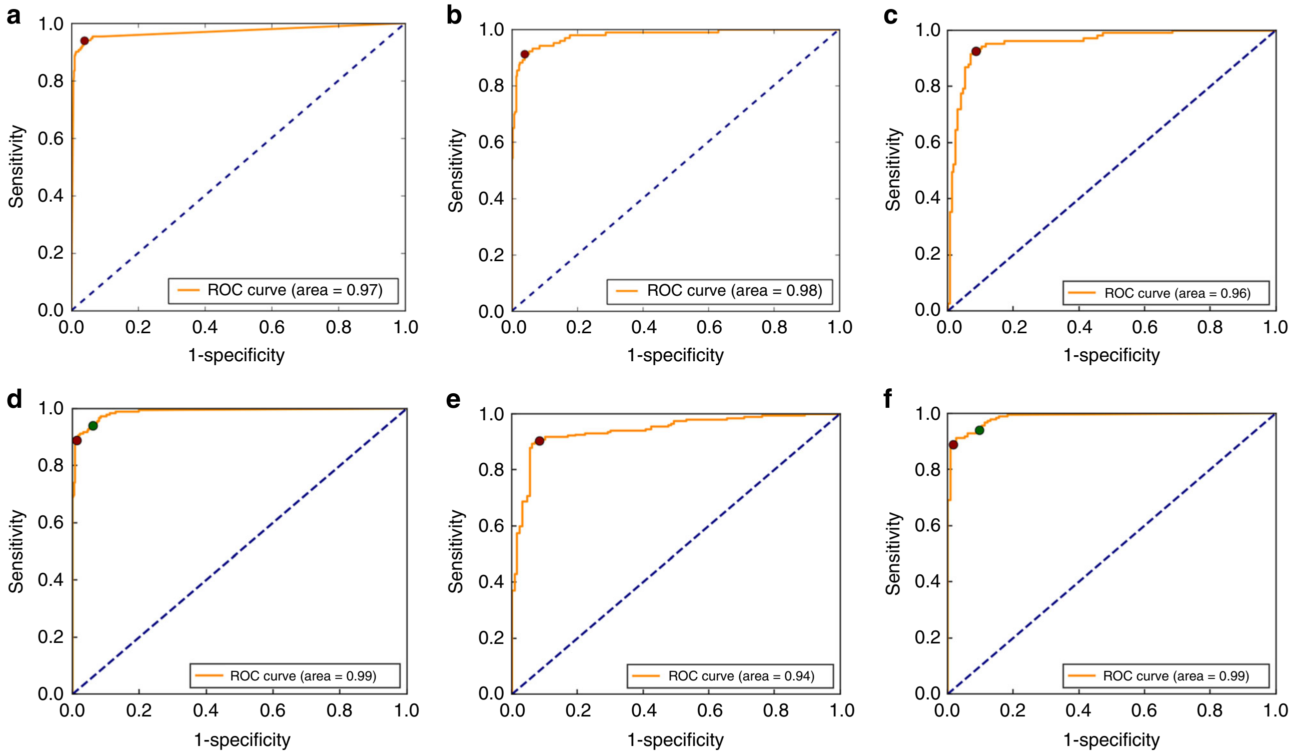


Fig. 4 Diagnostic receiver operating characteristics curves. Diagnostic receiver operating characteristics (ROC) curves, displaying the trade-offs between sensitivity and specificity for our narcolepsy biomarker for **a** training sample, **b** testing sample, **c** replication sample and **e** high pretest sample. **d**–**f** Adding HLA to model vastly increases specificity. Cut-off thresholds are presented for models with (red dot) and without HLA (green dot)

In the training data, a sensitivity of 94% and specificity of 96% was achieved, and in the testing data a sensitivity of 91% and specificity of 96% was achieved, while the sensitivity and specificity for the replication sample was 93 and 91%, respectively. When human leukocyte antigen (HLA) was added to this model (Fig. 4d–f), the sensitivity became 90% and the specificity rose to 99%, and an updated cut-off threshold of −0.53 was determined (green dot, Fig. 4d–f). Furthermore, in the high pretest sample we obtained a sensitivity and specificity of 90 and 92%, which rose to 90 and 98% when adding HLA. More descriptive statistics including 95% confidence intervals are found in Supplementary Table 6.

Discussion

In recent years, machine learning has been used to solve similar or more complex problems, such as labeling images, understanding speech and translating language, and have seen advancement to the point where humans are now sometimes outperformed[21–23], while also showing promising results in various medical fields[24–29]. Automatic classification of sleep stages using automatic algorithms is not novel[44,45], but only recently has this type of machine learning been applied and the effectiveness has only been demonstrated in a small numbers of sleep studies[46–49]. Because PSGs contain large amounts of manually annotated "gold standard" data, we hypothesized this method would be ideal to automatize sleep scoring. We have shown that machine learning can be used to score sleep stages in PSGs with high accuracy in multiple physical locations in various recording environments, using different protocols and hardware/software configurations, and in subjects with and without various sleep disorders.

After testing various machine learning algorithms with and without memory and specific encodings, we found increased robustness using a consensus of multiple algorithms in our prediction. The main reason for this is likely the sensitivity of each algorithm to particular aspects of each individual recording, resulting in increased or decreased predictability. Supplementary Figure 1b displays the correlations between different models. Models that incorporate an ensemble of different models generally have a higher overall correlation coefficient than singular models, and since individual models achieve similar performances, it stands to reason that these would achieve the highest performance. One potential source for this variability was, in addition to the stochastic nature of the training, the fact recordings were conducted in different laboratories that were using different hardware and filters, and had PSGs scored by technicians of various abilities. Another contributor was the presence of sleep pathologies in the dataset that could influence machine learning. Of the pathologies tested, only narcolepsy had a very significant effect on the correspondence between manual and machine learning methods ($p = 1.77 \times 10^{-15}$ vs $p = 7.53 \times 10^{-4}$ for sleep apnea for example) (Supplementary Tables 4 and 7). This was not surprising as the pathology is characterized by unusual sleep stage transitions, for example, transitions from wake to REM sleep, which may make human or machine learning staging more difficult. This result suggests that reporting inter-model variations in accuracy for each specific patient has value in flagging unusual sleep pathologies, so this metric is also reported by our detector.

Unlike previous attempts using automatic detector validations, we were able to include 70 subjects scored by 6 technicians in different laboratories (the IS-RC cohort)[31] to independently validate our best automatic scoring consensus algorithm. This allowed us to estimate the performance at 87% in comparison to the performance of a consensus score for every epoch among six expert technicians (ultimate gold standard) (Table 1). Including more scorers produces a better gold standard, and as Fig. 1a indicates, the model accuracy also increases with more scorers. Naturally, extrapolating from this should be done with caution; however, it is reasonable to assume that the accuracy would continue to increase with increased scorers. In comparison, performance of any individual scorer ranges from 74 to 85% when compared to the same six-scorer gold standard, keeping in mind this performance is artificially inflated since the same scorers evaluated are included in the gold standard (unbiased performance of any scorer versus consensus of remaining 5 scorers range from 69 to 80%. The best model achieves 87% accuracy using 5 scorers (Fig. 1a and Table 1), and is statistically higher than all scorers. As with human scorers, the biggest discrepancies in machine learning determination of sleep stages occurred between wake versus N1, N1 versus N2 and N2 versus N3. This is logical as these particular sleep stage transitions are part of a continuum, artificially defined and subjective. To give an example: an epoch comprised of 18% slow wave activity is considered N2 while an epoch comprised of 20% slow wave activity qualifies as N3. Overall, data indicate that our machine learning algorithm performs better than individual scorers, as typically used in clinical practice, or similar to the best of 5 scorers in comparison to a combination of 5 experts scoring each epoch by consensus. It is also able to score at higher resolution, i.e., 5 s, making it unnecessary to score sleep stages by 30 s epochs, an outdated rule dating from the time sleep was scored on paper. Although the data sample used for multi-scorer validation contained only female subjects, the scoring accuracy of our model was not seen to be affected by gender (Supplementary Table 3) in another analysis.

Using our models, and considering how typical T1N behaved in our sleep stage machine learning routines, we extracted features that could be useful to diagnose this condition. T1N is characterized by the loss of hypocretin-producing cells in the hypothalamus[3] and can be best diagnosed by measuring hypocretin levels in the CSF[11], a procedure that requires a lumbar puncture, a rarely performed procedure in the United States. At the symptomatic level, T1N is characterized by sleepiness, cataplexy (episodes of muscle weakness during wakefulness triggered by emotions) and numerous symptoms reflecting poor nocturnal sleep (insomnia) and symptoms of “dissociated REM sleep”. Dissociated REM sleep is reflected by the presence of unusual states of consciousness where REM sleep is intermingled with wakefulness, producing disturbing reports of dreams that interrupt wakefulness and seem real (dream-like hallucinations), or episodes where the sleeper is awake but paralyzed as in normal REM sleep (sleep paralysis). The current gold standard for T1N diagnosis is the presence of cataplexy and a positive MSLT. In a recent large study of the MSLT, specificity and sensitivity for T1N was 98.6% and 92.9% in comparing T1N versus controls, and 71.2% and 93.4% in comparing T1N versus other hypersomnia cases (high pretest probability cohort)[10].

Table 4 and Supplementary Table 5 reveal features found in nocturnal PSGs that discriminate type 1 narcoleptics and non-narcoleptics. One of the most prominent features, short latency REM sleep, bears great resemblance to the REM sleep latency, which is already used clinically to diagnose narcolepsy, although in this case it is calculated using fuzzy logic and thus represent a latency where accumulated sleep is suggestive of a high probability of REM sleep having occurred (as opposed to a discrete REM latency scored by a technician). A short REM latency during nocturnal PSG (typically 15 min) has recently been shown to be extremely specific (99%) and moderately sensitive (40–50%) for T1N[10,50]. The remaining selected features also describe a generally altered sleep architecture, particularly between REM sleep, light sleep and wake, aspects of narcolepsy already known and thus reinforcing their validity as biomarkers.

For example, the primary feature as determined by the RFE algorithm was the time taken until 5% of the accumulated sum of the probability products between stages W, N2 and REM had been reached (see also Table 4), which reflects the uncertainty between wakefulness, REM and N2 sleep at the beginning of the night. Specifically, for the n^{th} epoch, the model will output probabilities for each sleep stage, and the proto-feature $\boldsymbol{\Phi}_n$ is calculated as

$$\boldsymbol{\Phi}_n = p(W)\times p(\mathrm{N2}) + p(W)\times p(\mathrm{REM}) + p(\mathrm{N2})\times p(\mathrm{REM}). \tag{4}$$

The feature value is then calculated as the time it takes in minutes for the accumulated sum of $\boldsymbol{\Phi}_n$ to reach 5% of the total sum $\sum_n \boldsymbol{\Phi}_n$. Since each of probability product in $\boldsymbol{\Phi}_n$ reflects the staging uncertainty between each sleep stage pair, $\boldsymbol{\Phi}_n$ alone reflects the general sleep stage uncertainty for that specific epoch as predicted by the model. A very high value will be attained for epoch n if the probabilities for N2, W and REM are equally probable with probabilities for the remaining sleep stages being low or close to zero. A PSG with a high staging uncertainty between sleep and wake early in the night would reach the 5% threshold rapidly.

Using these features, we were able to determine an optimal cut-off that discriminated narcolepsy from controls and any other patients with as high specificity and sensitivity as the MSLT (Supplementary Table 6), notably when HLA typing is added. This is true for both the test and the never seen replication samples. Although we do observe a small drop in specificity in the replication sample, the efficacy of the detector was also tested in the context of naive patients with hypersomnia (high pretest probability sample), and performance found to be similar to the MSLT.

MSLT testing requires that patients spend an entire night and day in a sleep laboratory. The use of this novel biomarker could reduce time spent to a standard 8 h night recording, as done for the screening of other sleep pathologies (e.g., OSA), allowing improved recognition of T1N cases at a fraction of the cost. A positive predictive value could also be provided depending on the nature of the sample and known narcolepsy prevalence (low in general population screening, intermediary in overall clinic population sample and high in hypersomnia cohorts). It also opens the possibility of using home sleep recordings for diagnosing narcolepsy. In this direction, because of the probabilistic and automatic nature of our biomarker, estimates from more than one night could be automatically analyzed and combined over time, ensuring improved prediction. However, it is important to note that this algorithm will not replace the MSLT in the ability to predict excessive daytime sleepiness through the measure of mean sleep latency across daytime naps, which is an important characteristic of other hypersomnias.

In conclusion, models which classify sleep by assigning a membership function to each of five different stages of sleep for each analyzed segment were produced, and factors contributing to the performance were analyzed. The models were evaluated on different cohorts, one of which contained 70 subjects scored by 6 different sleep scoring technicians, allowing for inter-scorer reliability assessments. The most successful model, consisting of an ensemble of different models, achieved an accuracy of 87% on this dataset, and was statistically better performing than any individual scorer. It was also able to score sleep stages with high accuracy at lower time resolution (5 s), rendering the need for scoring per 30 s epoch obsolete. When predictions were weighted by the scorer agreement, performance rose to 95%, indicating a high consensus between the model and human scorers in areas of high scorer agreement. A final implementation was made using an ensemble with small variations of the best single model. This allowed for better predictions, while also providing a measure of uncertainty in an estimate.

When the staging data were presented as hypnodensity distributions, the model conveyed more information about the subject than through a hypnogram alone. This led to the creation of a biomarker for narcolepsy that achieved similar performance to the current clinical gold standard, the MSLT, but only requires a single sleep study. If increased specificity is needed, for example, in large-scale screening, HLA or additional genetic typing brings specificity above 99% without loss of sensitivity. This presents an option for robust, consistent, inexpensive and simpler diagnosis of subjects who may have narcolepsy, as such tests may also be carried out in a home environment.

This study shows how hypnodensity graphs can be created automatically from raw sleep study data, and how the resulting interpretable features can be used to generate a diagnosis probability for T1N. Another approach would be to classify narcolepsy directly from the neural network by optimizing the performance not only for sleep staging, but also for direct diagnosis by adding an additional softmax output, thereby creating a multitask classifier. This approach could lead to better predictions, since features are not then limited to by a designer imagination. A drawback of this approach is that features would no longer be as interpretable and meaningful to clinicians. If meaning could be extracted from these neural network generated features, this might open the door to a single universal sleep analysis model, covering multiple diseases. Development of such a model would require adding more subjects with narcolepsy and other conditions to the pool of training data.

Methods

Datasets. The success of machine learning depends on the size and quality of the data on which the model is trained and evaluated[51,52]. We used a large dataset comprised of several thousand sleep studies to train, validate and test/replicate our models. To ensure significant heterogeneity, data came from 10 different cohorts recorded at 12 sleep centers across 3 continents: SSC[10,32], WSC[32,33], IS-RC[31], JCTS[43], KHC[10,34], AHC[35], IHC[41], DHC[53], FHC and CNC[12]. Institutional review boards approved the study and informed consent was obtained from all participants. Technicians trained in sleep scoring manually labeled all sleep studies. Figure 5a–c summarizes the overall design of the study for sleep stage scoring and narcolepsy biomarker development. Supplementary Table 1 provides a summary of the size of each cohort and how it was used. In the narcolepsy biomarker aspect of the study, PSGs from T1N and other patients were split across most datasets to ensure heterogeneity in both the training and testing datasets. For this analysis, a few recordings with poor quality sleep studies, i.e., missing critical channels, with additional sensors or with a too short sleep duration (≤2 h) were excluded. A "never seen" subset cohort that included French and Chinese subjects (FHC and CNC) was also tested. Below is a brief description of each dataset.

Population-based Wisconsin Sleep Cohort. This cohort is a longitudinal study of state agency employees aged 37–82 years from Wisconsin, and it approximates a population-based sample (see Supplementary Table 1 for age at study) except for

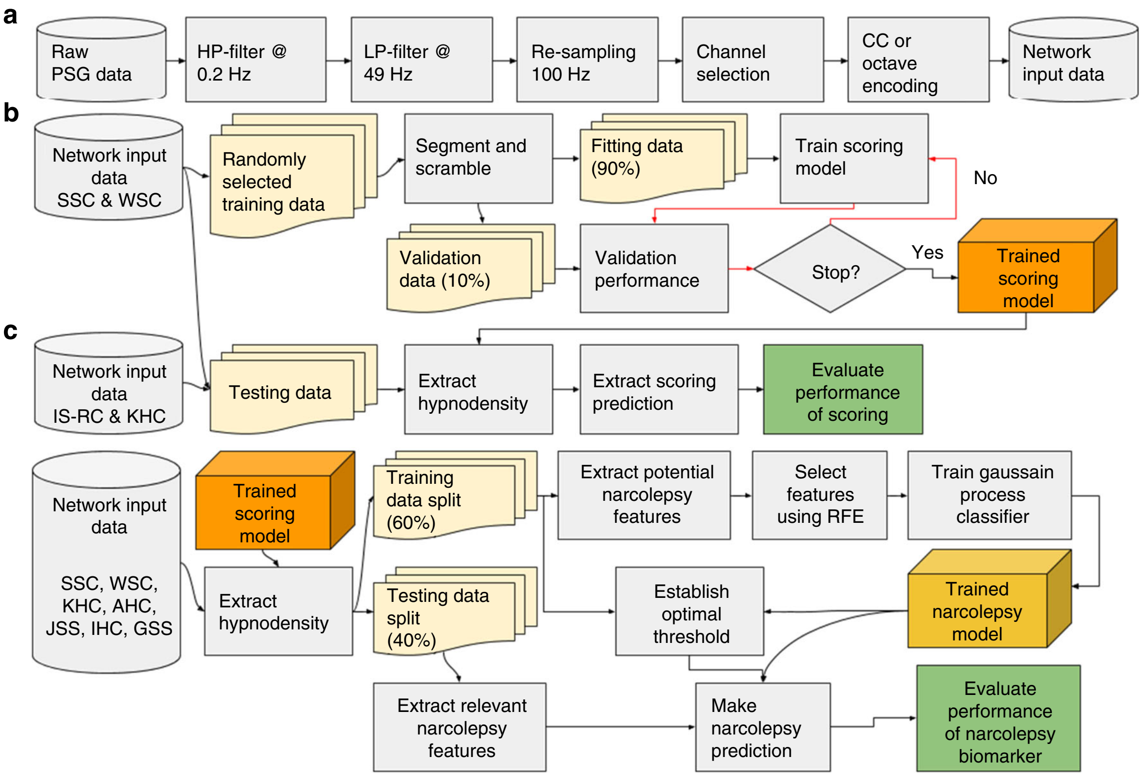

a
Raw PSG data
HP-filter @ 0.2 Hz
LP-filter @ 49 Hz
Re-sampling 100 Hz
Channel selection
CC or octave encoding
Network input data
b
Network input data SSC & WSC
Randomly selected training data
Segment and scramble
Fitting data (90%)
Train scoring model
No
Validation data (10%)
Validation performance
Stop?
Yes
Trained scoring model
c
Network input data IS-RC & KHC
Testing data
Extract hypnodensity
Extract scoring prediction
Evaluate performance of scoring
Network input data
SSC, WSC, KHC, AHC, JSS, IHC, GSS
Trained scoring model
Training data split (60%)
Extract potential narcolepsy features
Select features using RFE
Train gaussain process classifier
Extract hypnodensity
Testing data split (40%)
Establish optimal threshold
Trained narcolepsy model
Extract relevant narcolepsy features
Make narcolepsy prediction
Evaluate performance of narcolepsy biomarker

Fig. 5 Overall design of the study. **a** Pre-processing steps taken to achieve the format of data as it is used in the neural networks. One of the 5 channels is first high-pass filtered with a cut-off at 0.2 Hz, then low-pass filtered with a cut-off at 49 Hz followed by a re-sampling to 100 Hz to ensure data homogeneity. In the case of EEG signals, a channel selection is employed to choose the channel with the least noise. The data are then encoded using either the CC or the octave encoding. **b** Steps taken to produce and test the automatic scoring algorithm. A part of the SSC[10, 32] and WSC[32, 33] is randomly selected, as described in Supplementary Table 1. These data are then segmented in 5 min segments and scrambled with segments from other subjects to increase batch similarity during training. A neural network is then trained until convergence (evaluated using a separate validation sample). Once trained, the networks are tested on a separate part of the SSC and WSC along with data from the IS-RC[31] and KHC[10, 34]. **c** Steps taken to produce and test the narcolepsy detector. Hypnodensities are extracted from data, as described in Supplementary Table 1. These data are separated into a training (60%) and a testing (40%) split. From the training split, 481 potentially relevant features, as described in Supplementary Table 9, are extracted from each hypnodensity. The prominent features are maintained using a recursive selection algorithm, and from these features a GP classifier is created. From the testing split, the same relevant features are extracted, and the GP classifier is evaluated

the fact they are generally more overweight[33]. The study is ongoing, and dates to 1988. The 2167 PSGs in 1086 subjects were used for training, while 286 randomly selected PSGs were used for validation testing of the sleep stage scoring algorithm and narcolepsy biomarker training. Approximately 25% of the population have an Apnea Hypopnea Index (AHI) above 15/h and 40% have a PLMI above 15/h. A detailed description of the sample can be found in Young et al.[33] and Moore et al.[32]. The sample does not contain any T1N patients, and the three subjects with possible T1N were removed[54].

Patient-based Stanford Sleep Cohort. PSGs from this cohort were recorded at the Stanford Sleep Clinic dating back to 1999, and represent sleep disorder patients aged 18–91 years visiting the clinic (see Supplementary Table 1 for age at study). The cohort contains thousands of PSG recordings, but for this study we used 894 diagnostic (no positive airway pressure) recordings in independent patients that have been used in prior studies[30]. This subset contains patients with a range of different diagnoses including: sleep disordered breathing (607), insomnia (141), REM sleep behavior disorder (4), restless legs syndrome (23), T1N (25), delayed sleep phase syndrome (14) and other conditions (39). Description of the subsample can be found in Andlauer et al.[10] and Moore et al.[32]. Approximately 30% of subjects have an AHI above 15/h, or a PLMI above 15/h. The 617 randomly selected subjects were used for training the neural networks, while 277 randomly selected PSGs were kept for validation testing of the sleep stage scoring algorithm. These 277 subjects were also used for training the narcolepsy biomarker algorithm. The sample contains PSGs of 25 independent untreated subjects with T1N (12 with low CSF hypocretin-1, the others with clear cataplexy). A total of 26 subjects were removed from the study—4 due to poor data quality, and the rest because of medication use.

Patient-based Korean Hypersomnia Cohort. The Korean Hypersomnia Cohort is a high pretest probability sample for narcolepsy. It includes 160 patients with a primary complaint of excessive daytime sleepiness (see Supplementary Table 1 for age at study). These PSGs were used for testing the sleep scoring algorithm and for training the narcolepsy biomarker algorithm. No data were used for training the sleep scoring algorithm. Detailed description of the sample can be found in Hong et al.[34] and Andlauer et al.[10]. The sample contains PSGs of 66 independent untreated subjects with T1N and clear cataplexy. Two subjects were removed from the narcolepsy biomarker study because of poor data quality.

Patient-based Austrian Hypersomnia Cohort. Patients in this cohort were examined at the Innsbruck Medical University in Austria as described in Frauscher et al.[35]. The AHC contains 118 PSGs in 86 high pretest probability patients for narcolepsy (see Supplementary Table 1 for details). The 42 patients (81 studies) are clear T1N with cataplexy cases, with all but 3 having a positive MSLT (these three subjects had a MSL >8 min but multiple SOREMPs). The rest of the sample has idiopathic hypersomnia and type 2 narcolepsy. Four patients have an AHI >15/h and 25 had a PLMI >15/h. Almost all subjects had two sleep recordings performed, which were kept together such that no two recordings from the same subject were split between training and testing partitions.

Patient-based Inter-scorer Reliability Cohort. As Rosenberg and Van Hout[16] have shown, variation between individual scorers can sometimes be large, leading to an imprecise gold standard. To quantify this, and to establish a more accurate gold standard, 10 scorers from 5 different institutions, University of Pennsylvania, St. Luke's Hospital, University of Wisconsin at Madison, Harvard University and Stanford University, analyzed the same 70 full-night PSGs. For this study, scoring data from University of Pennsylvania, St. Luke's and Stanford were used. All subjects are female (see Supplementary Table 1 for details). This allowed for a much more precise gold standard, and the inter-scorer reliability could be quantified for a dataset, which could also be examined by automatic scoring algorithms. Detailed description of the sample can be found in Kuna et al.[31] and Malhotra and Avidan[6]. The sample does not contain any T1N patients.

The Jazz Clinical Trial Sample. This sample includes 7 baseline sleep PSGs from 5 sites taken from a clinical trial study of sodium oxybate in narcolepsy (SXB15 with 45 sites in Canada, United States, and Switzerland) conducted by Orphan Medical, now named Jazz Pharmaceuticals. The few patients included are those with clear and frequent cataplexy (a requirement of the trial) who had no stimulant or antidepressant treatment at baseline[43]. All seven subjects in this sample were used exclusively for training the narcolepsy biomarker algorithm.

Patient-based Italian Hypersomnia Cohort. Patients in this high pretest probability cohort (see Supplementary Table 1 for demographics) were examined at the IRCCS, Istituto delle Scienze Neurologiche ASL di Bologna in Italy as described in Pizza et al.[41]. The IHC contains 70 T1N patients (58% male, 29.5 ± 1.9 years old), with either documented low CSF hypocretin levels (59 cases, all but 2 HLA-DQB1*06:02 positive) or clear cataplexy, positive MSLTs and HLA positivity (11 subjects). As non-T1N cases with unexplained daytime somnolence, the cohort includes 77 other patients: 19 with idiopathic hypersomnia, 7 with type 2 narcolepsy and normal CSF hypocretin-1, 48 with a subjective complaint of excessive daytime sleepiness not confirmed by MSLT and 3 with secondary hypersomnia. Subjects in this cohort were used for training ($n = 87$) and testing ($n = 61$) the narcolepsy biomarker algorithm.

Patient-based Danish Hypersomnia Cohort. Patients in this cohort were examined at the Rigshospitalet, Glostrup, Denmark, as described in Christensen et al.[53]. The DHC contains 79 PSGs in controls and patients (see Supplementary Table 1 for details). Based on PSG, multiple sleep latency test and cerebrospinal fluid hypocretin-1 measures, the cohort includes healthy controls (19 subjects), patients with other sleep disorders and excessive daytime sleepiness (20 patients with CSF hypocretin-1 ≥110 pg/ml), narcolepsy type 2 (22 patients with CSF hypocretin-1 ≥110 pg/ml), and T1N (28 patients with CSF hypocretin-1 ≤110 pg/ml). All 79 subjects in this cohort were used exclusively for training the narcolepsy biomarker algorithm.

Patient-based French Hypersomnia Cohort. This cohort consists of 122 individual PSGs recorded at the Sleep-Wake Disorders Center, Department of Neurology, Gui-de-Chauliac Hospital, CHU Montpellier, France (see Supplementary Table 1 for demographics). The FHC contains 63 subjects with T1N (all but two tested with CSF hypocretin-1 ≤110 pg/ml, five below 18 years old, 55 tested for HLA, all positive for HLA-DQB1*06:02) and 22 narcolepsy type 2 (19 with CSF hypocretin-1 >200 pg/ml, and three subjects with CSF hypocretin-1 between 110 and 200 pg/ml, three HLA positive). The remaining 36 subjects are controls (15 tested for HLA, two with DQB1*06:02) without other symptoms of hypersomnia. The FHC was used as data for the replication study of the narcolepsy biomarker algorithm.

Patient-based Chinese Narcolepsy Cohort. This cohort contains 199 individual PSGs recorded (see Supplementary Table 1 for demographics). The CNC contains 67 subjects diagnosed with T1N exhibiting clear-cut cataplexy (55 tested HLA-DQB1*06:02 positive), while the remaining 132 subjects are randomly selected population controls (15 HLA-DQB1*06:02 positive, 34 HLA negative, remaining unknown)[12]. Together with the FHC, the CNC was used as data for the replication study of the narcolepsy biomarker algorithm.

American Academy of Sleep Medicine Sleep Study. The AASM ISR dataset is composed of a single control sleep study of 150 30 s epochs that was scored by 5234 ± 14 experienced sleep technologists for quality control purposes. Design of this dataset is described in Rosenberg and Van Hout[16].

Data labels, scoring and fuzzy logic. Sleep stages were scored by PSG-trained technicians using established scoring rules, as described in the AASM Scoring Manual[7]. In doing so, technicians assign each epoch with a discrete value. With a probabilistic model, like the one proposed in this study, a relationship to one of the fuzzy sets is inferred based on thousands of training examples labeled by many different scoring technicians.

The hypnodensity graph refers to the probability distribution over each possible stage for each epoch, as seen in Fig. 2a, b. This allows more information to be conveyed, since every epoch of sleep within the same stage is not identical. For comparison with the gold standard, however, a discrete value must be assigned from the model output as:

$$\hat{y} = \operatorname{argmax}_{\mathbf{y}_i} \sum_{i}^{N} \mathbf{P}_i(\mathbf{y}_i|\mathbf{x}_i), \tag{5}$$

where $\mathbf{P}_i(\mathbf{y}_i|\mathbf{x}_i)$ is a vector with the estimated probabilities for each sleep stage in the i^{th} segment, N is the number of segments an epoch is divided into and $\hat{y}$ is the estimated label.

Sleep scoring technicians score sleep in 30 s epochs, based on what stage they assess is represented in the majority of the epoch—a relic of when recordings were done on paper. This means that when multiple sleep stages are represented, more than half of the epoch may not match the assigned label. This is evident in the fact that the label accuracy decreases near transition epochs[20]. One solution to this problem is to remove transitional regions to purify each class. However, this has the disadvantage of under-sampling transitional stages, such as N1, and removes the context of quickly changing stages, as is found in a sudden arousal. It has been demonstrated that the negative effects of imperfect "noisy" labels may be mitigated if a large enough training dataset is incorporated and the model is robust to overfitting[41]. This also assumes that the noise is randomly distributed with an accurate mean—a bias cannot be canceled out, regardless of the amount of training data. For these reasons, all data including those containing sleep transitions were included. Biases were evaluated by incorporating data from several different scoring experts cohorts and types of subjects.

To ensure quick convergence, while also allowing for long-term dependencies in memory-based models, the data were broken up in 5 min blocks and shuffled to minimize the shift in covariates during training caused by differences between subjects. To quantify the importance of segment sizes, both 5 s and 15 s windows were also tested.

Data selection and pre-processing. A full-night PSG involves recording many different channels, some of which are not necessary for sleep scoring[55]. In this study, EEG, C3 or C4, and O1 or O2, chin EMG and the left and right EOG channels were used, with reference to the contralateral mastoid. Poor electrode connections are common when performing a PSG analysis. This can lead to a noisy recording, rendering it useless. To determine whether right or left EEG channels were used, the noise of each was quantified by dividing the EEG data in 5 min segments, and extracting the Hjorth parameters[56]. These were then log-transformed, averaged and compared with a previously established multivariate distribution, based on the WSC[32,33] and SSC[10,32] training data. The channel with lowest Mahalanobis distance[57] to this distribution was selected. The log transformation has the advantage of making flat signals/disconnects as uncommon as very noisy signals, in turn making them less likely to be selected. To minimize heterogeneity across recordings, and at the same time reducing the size of the data, all channels were down-sampled to 100 Hz. Additionally, all channels were filtered with a fifth-order two-direction infinite impulse response (IIR) high-pass filter with cut-off frequency of 0.2 Hz and a fifth-order two-direction IIR low-pass filter with cut-off frequency of 49 Hz. The EMG signal contains frequencies well above 49 Hz, but since much data had been down-sampled to 100 Hz in the WSC, this cut-off was selected for all cohorts. All steps of the pre-processing are illustrated in Fig. 5a.

Convolutional and recurrent neural networks. Convolutional neural networks (CNNs) are a class of deep learning models first developed to solve computer vision problems[30]. A CNN is a supervised classification model in which a low level, such as an image, is transformed through a network of filters and sub-sampling layers. Each layer of filters produces a set of features from the previous layer, and as more layers are stacked, more complex features are generated. This network is coupled with a general-purpose learning algorithm, resulting in features produced by the model reflecting latent properties of the data rather than the imagination of the designer. This property places fewer constrictions on the model by allowing more flexibility, and hence the predictive power of the model will increase as more data are observed. This is facilitated by the large number of parameters in such a model, but may also necessitate a large amount of training data. Sleep stage scoring involves a classification of a discrete time series, in which adjacent segments are correlated. Models that incorporate memory may take advantage of this and may lead to better overall performance by evening out fluctuations. However, these fluctuations may be the defining trait or anomaly of some underlying pathology (such as narcolepsy, a pathology well known to involve abnormal sleep stages transitions), present in only a fraction of subjects, and perhaps absent in the training data. This can be thought of similarly to a person with a speech impediment: the contextual information will ease the understanding, but knowing only the output, this might also hide the fact that the person has such a speech impediment. To analyze the importance of this, models with and without memory were analyzed. Memory can be added to such a model by introducing recurrent connections in the final layers of the model. This turns the model into a recurrent neural network (RNN). Classical RNNs had the problem of vanishing or exploding gradients, which meant that optimization was very difficult. This problem was solved by changing the configuration of the simple hidden node into a LSTM cell[58]. Models without this memory are referred to as FF models. A more in-depth explanation of CNNs including application areas can be found in the review article on deep learning by LeCun et al.[30] and the deep learning textbook by Goodfellow et al.[59]. For a more general introduction to machine learning concepts, see the textbook by Bishop[60].

Data input and transformations. Biophysical signals, such as those found in a PSG, inherently have a low signal to noise ratio, the degree of which varies between subjects, and hence learning robust features from these signals may be difficult. To circumvent this, two representations of the data that could minimize these effects were selected. An example of each decomposition is shown in Fig. 6a.

Octave encoding maintains all information in the signal, and enriches it by repeatedly removing the top half of the bandwidth (i.e., cut-off frequencies of 49, 25, 12.5, 6.25 and 3.125 Hz) using a series of low-pass filters, yielding a total of 5 new channels for each original channel. At no point is a high-pass filter applied. Instead, the high frequency information may be obtained by subtracting lower frequency channels—an association the neural networks can make, given their universal approximator properties[61]. After filtration, each new channel is scaled to the 95th percentile and log modulus transformed:

$$\mathbf{x}_{\mathrm{scaled}} = \mathrm{sign}(\mathbf{x}) \cdot \log\left(\frac{|\mathbf{x}|}{P_{95}(\mathbf{x})} + 1\right). \tag{6}$$

The initial scaling places 95% of the data between −1 and 1, a range in which the log modulus is close to linear. Very large values, such as those found in particularly noisy areas, are attenuated greatly. Some recordings are noisy, making the 95th percentile significantly higher than what the physiology reflects. Therefore, instead of selecting the 95th percentile from the entire recording, the recording is separated into 50% overlapping 90 min segments, from which the 95th percentile is extracted. The mode of these values is then used as a scaling reference. In general, scaling and normalization is important to ensure quick convergence as well as generalization in neural networks. The decomposition is done in the same way on every channel, resulting in 25 new channels in total.

CC encoding, using a CC function, underlying periodicities in the data are revealed while noise is attenuated. White noise is by definition uncorrelated; its autocorrelation function is zero everywhere except lag zero. It is this property that is utilized, even though noise cannot always be modeled as such. PSG signals are often obscured by undesired noise that is uncorrelated with other aspects of the signals. An example CC between a signal segment and an extended version of the same signal segment is shown in Supplementary Figure 5. Choosing the CC in this manner over a standard autocorrelation function serves two purposes: the slow frequencies are expressed better, since there is always full overlap between the two signals (some of this can be adjusted with the normal autocorrelation function using an unbiased estimate); and the change in fluctuations over time within a segment is expressed, making the function reflect aspects of stationarity. Because this is the CC between a signal and an extended version of itself, the zero lag represents the power of that segment, as is the case in an autocorrelation function.

Frequency content with a time resolution may also be expressed using time-frequency decompositions, such as spectrograms or scalograms; however, one of the key properties of a CNN is the ability to detect distinct features anywhere in an input, given its property of equivariance[62]. A CC function reveals an underlying set of frequencies as an oscillation pattern, as opposed to a spectrogram, where frequencies are displayed as small streaks or spots in specific locations, corresponding to frequencies at specific times. The length and size of each CC reflects the expected frequency content and the limit of quasi-stationarity (i.e., how quickly the frequency content is expected to change).

The EOG signal reveals information about eye movements such as REMs, and to some extent EEG activity[6,7]. In the case of the EOG signal, the relative phase between the two channels is of great importance to determine synchronized eye movements, and hence a CC of opposite channels (i.e., either the extended or zero padded signal is replaced with the opposite channel) is also included. The slowest eye movements happen over the course of several seconds[6,7], and hence a segment length of 4 s was selected for the correlation functions. To maintain resolution flexibility with the EEG, an overlap of 3.75 s was chosen.

In the case of the EMG signal, the main concern is the signal amplitude and the temporal resolution, not the actual frequencies. As no relevant low-frequency content is expected, a segment length of 0.4 s and an overlap of 0.25 s was selected.

As with the octave encoding, the data are scaled, although only within segments:

$$D_i = \frac{\gamma_{\mathbf{x}_i\mathbf{y}_i} \cdot \log\left(1 + \max\left(\left|\gamma_{\mathbf{x}_i\mathbf{y}_i}\right|\right)\right)}{\max\left(\left|\gamma_{\mathbf{x}_i\mathbf{y}_i}\right|\right)}, \tag{7}$$

where D_i is the scaled correlation function and $\gamma_{\mathbf{x}_i\mathbf{y}_i}$ is the unscaled correlation function.

Architectures of applied CNN models. The architecture of a CNN typically reflects the complexity of the problem that is being solved and how much training data are available, as a complex model has more parameters than a simple model, and is therefore more likely to over-fit. However, much of this may be solved using proper regularization. Another restriction is the resources required to train a model—deep and complex models require far more operations and will therefore take longer to train and operate. In this study, no exhaustive hyper-parameter optimization was carried out. The applied architectures were chosen on the basis of other published models[63]. Since the models utilized three separate modalities (EEG, EOG and EMG), three separate sub-networks were constructed. These were followed by fully connected layers combining the inputs from each sub-network, which were passed onto a softmax output (Fig. 6b, Supplementary Figure 3). Models that utilize memory have fully connected hidden units replaced with LSTM cells and recurrent connections added between successive segments. Networks of two different sizes are evaluated to quantify the effect of increasing complexity.

Training of CNN models. Training the models involves optimizing parameters to minimize a loss function evaluated across a training dataset. The loss function was defined as the cross-entropy with L2 regularization:

$$L(\boldsymbol{\omega}) = \frac{1}{N}\sum\nolimits_{i=1}^{N} H\left(\mathbf{y}_i, \widehat{\mathbf{y}}_i\right) + L2 = \frac{1}{N}\sum\nolimits_{i=1}^{N} \mathbf{y}_i\log\widehat{\mathbf{y}}_{\boldsymbol{i}} + \left(1 - \mathbf{y}_i\right)\log(1 - \widehat{\mathbf{y}}_{\boldsymbol{i}}) + \lambda\,||\boldsymbol{\omega}||_2^2, \tag{8}$$

where $\mathbf{y}_i$ is the true class label of the $\boldsymbol{i}^{\mathrm{th}}$ window, $\widehat{\mathbf{y}}_{\boldsymbol{i}}$ is the estimated probability of the $\boldsymbol{i}^{\mathrm{th}}$ window, $\boldsymbol{\omega}$ is the parameter to be updated and λ is the weight decay parameter set at 0.00001. The model parameters were initialized with $N(0, 0.01)$, and trained until convergence using stochastic gradient decent with momentum[64]. Weight updates were done as: $\boldsymbol{\omega}_{t+1} = \boldsymbol{\omega}_t + \eta\mathbf{v}_{t+1}$ with $\mathbf{v}_{t+1} = \alpha\mathbf{v}_t - \frac{\delta\mathrm{E}}{\delta\boldsymbol{\omega}_t}$ where α is the momentum set at 0.9, $\mathbf{v}_t$ is the learning velocity, initialized at 0, and η is the learning rate, initially set at 0.005. The learning rate was gradually reduced with an exponential decay $\eta = \eta_0 \cdot \mathrm{e}^{-t/\tau}$ where t is the number of updates and τ is a time constant, here set to 12,000.

Overfitting was avoided using a number of regularization techniques, including batch normalization[65], weight decay[66] and early stopping[67]. Early stopping is accomplished by scheduling validation after every 50th training batch. This is done by setting aside 10% of the training data. Training is stopped if the validation

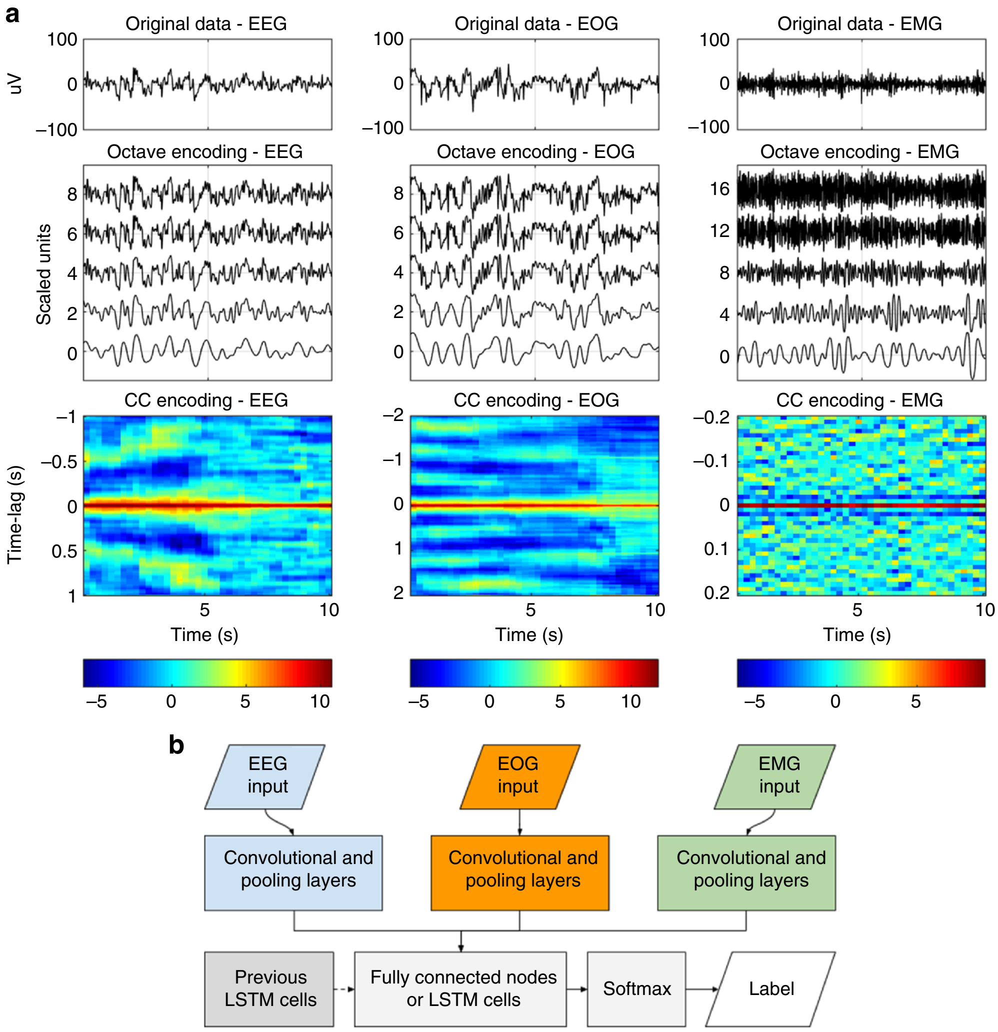


Fig. 6 Neural network strategy. **a** An example of the octave and the CC encoding on 10 s of EEG, EOG and EMG data. These processed data are fed into the neural networks in one of the two formats. The data in the octave encoding are offset for visualization purposes. Color scale is unitless. **b** Simplified network configuration, displaying how data are fed and processed through the networks. A more detailed description can be found in Supplementary Figure 3

accuracy starts to decrease, as a sign of overfitting. For LSTM networks, dropout[68] was included, set at 0.5 while training. This ensured that model parameters generalized to the validation data and beyond. During training, data batches were selected at random. Given the stochastic nature of the training procedure, it was likely that two realizations of the same model would not lead to the same results, since models end up in different local minima. To measure the effect of this, two realizations were made of each model.

Apart from model realizations, we also investigated the effect of ensembling our sleep stage classification model. In general, ensemble models can yield higher predictive performance than any single model by attacking a classification or regression problem from multiple angles. For our specific use case, this resolves into forming a sleep stage prediction based on the predictions of all the models in the given ensemble. We tested several ensembles containing various numbers of model architectures and data encodings, as described in Supplementary Table 8.

Performance comparisons of generated CNN models. As stated, the influences of many different factors were analyzed. These included: using octave or CC encoding, short (5 s) or long (15 s) segment lengths, low or high complexity, with or without LSTM, and using a single or two realizations of a model. To quantify the effect of each, a 2^5-factorial experiment was designed. This led to 32 different models (Supplementary Table 8). Comparison between models was done on a per-epoch basis.

Hypnodensity as feature for the diagnosis of T1N. To quantify narcolepsy-like behavior for a single recording i, features were generated based on a proto-feature derived from k-combinations of $\mathcal{S} = \{\mathrm{W}, \mathrm{REM}, \mathrm{N1}, \mathrm{N2}, \mathrm{N3}\}$. For the nth 5, 15 or 30 s segment in recording i, we take a single k-combination in the set of all k-combinations, and calculate the proto-feature as the sum of the pair-wise products of the elements in the single k-combination, such that

$$\Phi_n^{(i)}(\mathcal{S}_k) = \sum_{\zeta\in[\mathcal{S}_k]^2} \prod_{s\in\zeta} p\left(s|\mathbf{x}_n^{(i)}\right), \quad p\in[0,1], \qquad (9)$$

where $\Phi_n^{(i)}$ is the proto-feature for the n^{th} segment in recording i, $\zeta \in [\mathcal{S}_k]^2$ is a 2-tuple, or pair-wise combination, in the set of all pair-wise combinations in the k-combination of $\mathcal{S}$ and s is a single element, or sleep stage, in ζ. For $k = 1, \ldots, 5$, there exist 31 different $\mathcal{S}_k$, e.g., {Wake, REM}, {N1, N2, N3} etc., as shown in Supplementary Table 9. $p\left(s|\mathbf{x}_n^{(i)}\right)$ is the predicted probability of a 5, 15 or 30 s epoch belonging to a certain class in $\mathcal{S}$, given the data $\mathbf{x}_n^{(i)}$. For every value of k, 15 features based on the mean, derivative, entropy and cumulative sum were extracted, as shown in Supplementary Table 10.

Additional features for T1N diagnosis. In addition to above, another set of features reflecting abnormal sleep stage sequencing in T1N was investigated.

One set of such features was selected because they have been found to differentiate T1N from other subjects in prior studies[37,69–72]. These include: nocturnal sleep REM latency (REML)[10], presence of a nightly SOREMP (REML ≤15 min)[10], presence and number of SOREMPs during the night (SOREMPs defined as REM sleep occurring after at least 2.5 min of wake or stage 1) and nocturnal sleep latency (a short sleep latency is common in narcolepsy)[37]. Other features include a NREM Fragmentation index described in Christensen et al.[37]. (N2 and N3 combined to represent unambiguous NREM and N1 and wake combined to denote wake, NREM fragmentation defined as 22 or more occurrences where sustained N2/N3 (90 s) is broken by at least 1 min of N1/Wake), and the number of W/N1 hypnogram bouts as defined by Christensen et al.[37]. (N1 and wake combined to indicate wakefulness and a long period defined as 3 min or more). In this study we also explore: the cumulative wake/N1 duration for wakefulness periods shorter than 15 min; cumulative REM duration following wake/N1 periods longer than 2.5 min; and total nightly SOREMP duration defined as the sum of REM epochs following 2.5 min W/N1 periods.

Another set of 9 features reflecting hypnodensity sleep stage distribution was also created as follows. As noted in Supplementary Figure 4, stages of sleep accumulate, forming peaks. These peaks were then used to create 9 new features based on the order of the peaks, expressing a type of transition (W to N2, W to REM, REM to N3 etc.). If the height of the n^{th} peak is denoted as φ_n, the transition value τ is calculated as the geometric mean between successive peaks:

$$\tau_n = \sqrt{\varphi_n \cdot \varphi_{n+1}}. \qquad (10)$$

Due to their likeness, W and N1 peaks were added to form a single type. All transitions of a certain type were added together to form a single feature. A lower limit of 10 was imposed on peaks to avoid spurious peaks. If two peaks of the same type appeared in succession the values were combined into a single peak.

Gaussian process models for narcolepsy diagnosis. To avoid overfitting, and at the same time produce interpretable results, a RFE algorithm was employed, as described in Guyon et al.[42]. Post screening, the most optimal features ($n = 38$) were used in a GP classifier as described below. GP classifiers are non-parametric probabilistic models that produce robust non-linear decision boundaries using kernels, and unlike many other classification tools, provide an estimate of the uncertainty. This is useful when combining estimates, but also when making a diagnosis; if an estimate is particularly uncertain, a doctor may opt for more tests to increase certainty before making a diagnosis. In a GP, a training dataset is used to optimize a set of hyper-parameters, which specify the kernel function, the basis function coefficients, here a constant, noise variance, and to form the underlying covariance and mean function from which inference about new cases are made[73]. In this case, the kernel is the squared exponential: $\sigma_f^2 \exp\left[\frac{-|\mathbf{x}-\mathbf{x}'|^2}{2l^2}\right]$. Two classes were established: narcolepsy type 1 and “other”, which contains every other subject. These were labeled 1 and −1 respectively, placing all estimates in this range. For more information on GP in general, see the textbook by Rasmussen and Williams[73], while more information on variational inference for scalable GP classification can be found in the paper by Hensman et al.[74] and Matthews et al.[75].

HLA-DQB1*06:02 testing. HLA testing plays a role in T1N diagnosis, as 97% of patients are DQB1*06:02 positive when the disease is defined biochemically by low CSF hypocretin-1[5] or by the presence of cataplexy and clear MSLT findings[10]. As testing for HLA-DQB1*06:02 only requires a single blood test, models in which this feature was included were also tested. The specific feature was implemented as a binary-valued predictor, resulting in negative narcolepsy predictions for subjects with a negative HLA test result.

High pretest probability sample. MSLTs are typically performed in patients with daytime sleepiness that cannot be explained by OSA, insufficient/disturbed sleep or circadian disturbances. These patients have a higher pretest probability of having T1N than random clinical patients. Patients are then diagnosed with type 1 or type 2 narcolepsy, idiopathic hypersomnia or subjective sleepiness based on MSLT results, cataplexy symptoms and HLA results (if available). To test whether our detector differentiates T1N from these other cases with unexplained sleepiness, we conducted a post hoc analysis of the detector performance in these subjects extracted from both the test and replication datasets.

Data availability

All the software is made available in GitHub at: https://github.com/stanford-stages/stanford-stages. We asked all contributing co-authors whether we could make the anonymized EDF available, together with age, sex and T1N diagnosis (Y/N). The SSC[10,32] (E.M.), the IS-RC[31] (S.T.K., C.K., P.K.S.), the KHC (S.C.H.), the HIS (G. P.), the DHS (P.J.), the FHC (Y.D.) and associated data are available at https://stanfordmedicine.app.box.com/s/r9e92ygq0erf7hn5re6j51aaggf50jly. The AHC (B. H.) and the CNC[12] (F.H.) are available from the corresponding investigator on reasonable request. The WSC[32,33] data analyzed during the current study are not publicly available due to specific language contained in informed consent documents limiting use of WSC human subjects' data to specified institutionally approved investigations. However, WSC can be made available from P.E.P. on reasonable request and with relevant institutional review board(s) approval. The JCTS[43] and AASM ISR[16] dataset are available from the corresponding institutions on reasonable request.

Received: 1 October 2017 Accepted: 15 October 2018
Published online: 06 December 2018

References

1. Krieger, A. C. *Social and Economic Dimensions of Sleep Disorders, An Issue of Sleep Medicine Clinics* (Elsevier, Philadelphia, PA, 2007).
2. American Academy of Sleep Medicine. *International Classification of Sleep Disorders*, 3rd edn (American Academy of Sleep Medicine, Darien, IL, 2014).
3. Peyron, C. et al. A mutation in a case of early onset narcolepsy and a generalized absence of hypocretin peptides in human narcoleptic brains. *Nat. Med.* **6**, 991–997 (2000).
4. Kornum, B. R. et al. Narcolepsy. *Nat. Rev. Dis. Prim.* **3**, 16100 (2017).
5. Han, F. et al. HLA DQB1*06:02 negative narcolepsy with hypocretin/orexin deficiency. *Sleep* **37**, 1601–1608 (2014).
6. Malhotra, R. K. & Avidan, A. Y. Sleep Stages and Scoring Technique. In *Atlas of Sleep Medicine* (eds Chokroverty, S. & Thomas, R. J.), 77–99 (Elsevier, Philadelphia, PA, 2014).
7. Berry, R. B. et al. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, version 2.4* (American Academy of Sleep Medicine, Darien, IL, 2017).
8. Subramanian, S., Hesselbacher, S., Mattewal, A. & Surani, S. Gender and age influence the effects of slow-wave sleep on respiration in patients with obstructive sleep apnea. *Sleep Breath.* **17**, 51–56 (2013).
9. Littner, M. R. et al. Practice parameters for clinical use of the multiple sleep latency test and the maintenance of wakefulness test. *Sleep* **28**, 113–121 (2005).
10. Andlauer, O. et al. Nocturnal rapid eye movement sleep latency for identifying patients with narcolepsy/hypocretin deficiency. *JAMA Neurol.* **70**, 891–902 (2013).
11. Mignot, E. et al. The role of cerebrospinal fluid hypocretin measurement in the diagnosis of narcolepsy and other hypersomnias. *Arch. Neurol.* **59**, 1553–1562 (2002).
12. Andlauer, O. et al. Predictors of hypocretin (orexin) deficiency in narcolepsy without cataplexy. *Sleep* **35**, 1247–1255 (2012).
13. Luca, G. et al. Clinical, polysomnographic and genome-wide association analyses of narcolepsy with cataplexy: a European Narcolepsy Network study. *J. Sleep Res.* **22**, 482–495 (2013).
14. Dauvilliers, Y. et al. Effect of age on MSLT results in patients with narcolepsy-cataplexy. *Neurology* **62**, 46–50 (2004).
15. Moscovitch, A., Partinen, M. & Guilleminault, C. The positive diagnosis of narcolepsy and narcolepsy's borderland. *Neurology* **43**, 55–60 (1993).
16. Rosenberg, R. S. & Van Hout, S. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. *J. Clin. Sleep Med.* **9**, 81–87 (2013).
17. Zhang, X. et al. Process and outcome for international reliability in sleep scoring. *Sleep Breath.* **19**, 191–195 (2015).
18. Danker-Hopfe, H. et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *J. Sleep Res.* **18**, 74–84 (2009).
19. MacLean, A. W., Lue, F. & Moldofksy, H. The reliability of visual scoring of alpha EEG activity during sleep. *Sleep* **18**, 565–569 (1995).
20. Kim, Y., Kurachi, M., Horita, M., Matsuura, K. & Kamikawa, Y. Agreement of visual scoring of sleep stages among many laboratories in Japan: effect of a supplementary definition of slow wave on scoring of slow wave sleep. *J. Psychiatry Clin. Neurosci.* **47**, 91–97 (1993).
21. Hinton, G. et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Process. Mag.* **83**, 82–97 (2012).
22. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In *2015 IEEEInternational Conference on Computer Vision (ICCV)* 1026–1034 (IEEE, Santiago, Chile, 2015).
23. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, Las Vegas, NV, 2016).
24. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
25. Ting, D. S. W. et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* **318**, 2211–2223 (2017).
26. Bejnordi, B. E. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).
27. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
28. Cheng, J.-Z. et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. *Sci. Rep.* **6**, 24454 (2016).
29. Lakhani, P. & Sundaram, B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* **284**, 574–582 (2017).

30. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
31. Kuna, S. T. et al. Agreement in computer-assisted manual scoring of polysomnograms across sleep centers. *Sleep* **36**, 583–589 (2013).
32. Moore, H. I. et al. Design and validation of a periodic leg movement detector. *PLoS One* **9**, e114565 (2014).
33. Young, T. et al. Burden of sleep apnea: rationale, design, and major findings of the Wisconsin Sleep Cohort study. *WMJ* **108**, 246–249 (2009).
34. Hong, S. C. et al. A study of the diagnostic utility of HLA typing, CSF hypocretin-1 measurements, and MSLT testing for the diagnosis of narcolepsy in 163 Korean patients with unexplained excessive daytime sleepiness. *Sleep* **29**, 1429–1438 (2006).
35. Frauscher, B. et al. Delayed diagnosis, range of severity, and multiple sleep comorbidities: a clinical and polysomnographic analysis of 100 patients of the Innsbruck Narcolepsy Cohort. *J. Clin. Sleep Med.* **9**, 805–812 (2013).
36. Mander, B. A., Winer, J. R. & Walker, M. P. Sleep and human aging. *Neuron* **94**, 19–36 (2017).
37. Christensen, J. A. E. et al. Sleep-stage transitions during polysomnographic recordings as diagnostic features of type 1 narcolepsy. *Sleep Med.* **16**, 1558–1566 (2015).
38. Olsen, A. V. et al. Diagnostic value of sleep stage dissociation as visualized on a 2-dimensional sleep state space in human narcolepsy. *J. Neurosci. Methods* **282**, 9–19 (2017).
39. Jensen, J. B. et al. Sleep-wake transition in narcolepsy and healthy controls using a support vector machine. *J. Clin. Neurophysiol.* **31**, 397–401 (2014).
40. Vassalli, A. et al. Electroencephalogram paroxysmal theta characterizes cataplexy in mice and children. *Brain* **136**, 1592–1608 (2013).
41. Pizza, F. et al. Nocturnal sleep dynamics identify narcolepsy type 1. *Sleep* **38**, 1277–1284 (2015).
42. Guyon, I., Weston, J. & Barnhill, S. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002).
43. International Xyrem Study Group. A double-blind, placebo-controlled study demonstrates sodium oxybate is effective for the treatment of excessive daytime sleepiness in narcolepsy. *J. Clin. Sleep Med.* **1**, 391–397 (2005).
44. Anderer, P. et al. 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).
45. Olesen, A. N., Christensen, J. A. E., Sorensen, H. B. D. & Jennum, P. J. A noise-assisted data analysis method for automatic EOG-based sleep stage classification using ensemble learning. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 3769–3772 (IEEE, Orlando, FL, 2016).
46. Boostani, R., Karimzadeh, F. & Nami, M. A comparative review on sleep stage classification methods in patients and healthy individuals. *Comput. Methods Prog. Biomed.* **140**, 77–91 (2017).
47. Lajnef, T. et al. Learning machines and sleeping brains: automatic sleep stage classification using decision-tree multi-class support vector machines. *J. Neurosci. Methods* **250**, 94–105 (2015).
48. da Silveira, T. L. T., Kozakevicius, A. J. & Rodrigues, C. R. Single-channel EEG sleep stage classification based on a streamlined set of statistical features in wavelet domain. *Med. Biol. Eng. Comput.* **55**, 343–352 (2017).
49. Ronzhina, M. et al. Sleep scoring using artificial neural networks. *Sleep Med. Rev.* **16**, 251–263 (2012).
50. Reiter, J., Katz, E., Scammell, T. E. & Maski, K. Usefulness of a nocturnal SOREMP for diagnosing narcolepsy with cataplexy in a pediatric population. *Sleep* **38**, 859–865 (2015).
51. Banko, M. & Brill, E. Scaling to very very large corpora for natural language disambiguation. In *ACL '01 Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* 26–33 (Association for Computational Linguistics, Stroudsburg, PA, 2001).
52. Shotton, J. et al. Real-time human pose recognition in parts from single depth images. *Stud. Comput. Intell.* **411**, 119–135 (2013).
53. Christensen, J. A. E. et al. Novel method for evaluation of eye movements in patients with narcolepsy. *Sleep Med.* **33**, 171–180 (2017).
54. Goldbart, A. et al. Narcolepsy and predictors of positive MSLTs in the Wisconsin Sleep Cohort. *Sleep* **37**, 1043–1051 (2014).
55. Silber, M. H. et al. The visual scoring of sleep in adults. *J. Clin. Sleep Med.* **3**, 121–131 (2007).
56. Hjorth, B. EEG analysis based on time domain properties. *Electroencephalogr. Clin. Neurophysiol.* **29**, 306–310 (1970).
57. Mahalanobis, P. C. On the generalised distance in statistics. *Proc. Natl. Inst. Sci. India* **2**, 49–55 (1936).
58. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
59. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, Cambridge, 2016).
60. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, New York, 2006).
61. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **2**, 359–366 (1989).
62. Lenc, K. & Vedaldi, A. Understanding image representations by measuring their equivariance and equivalence. In *2015IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 991–999 (IEEE, Boston, MA, 2015).
63. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *2015 International Conference on Learning Representation (ICLR)* 1–14 (ICLR, San Diego, CA, 2015).
64. Polyak, B. T. Some methods of speeding up the convergence of iteration methods. *USSR Comput. Math. Math. Phys.* **4**, 1–17 (1964).
65. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. *Proc. 32nd Int. Conf. Mach. Learn. PLMR* **37**, 448–456 (2015).
66. Krogh, A. & Hertz, J. A. A simple weight decay can improve generalization. *Adv. Neural Inf. Process. Syst.* **4**, 950–957 (1992).
67. Caruana, R., Lawrence, S. & Giles, L. Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping. In *Proc. Advances in Neural Information Processing Systems 13* 402–408 (MIT Press, Cambridge, MA, 2001).
68. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
69. Roth, T. et al. Disrupted nighttime sleep in narcolepsy. *J. Clin. Sleep Med.* **9**, 955–965 (2013).
70. Hansen, M. H., Kornum, B. R. & Jennum, P. Sleep-wake stability in narcolepsy patients with normal, low and unmeasurable hypocretin levels. *Sleep Med.* **34**, 1–6 (2017).
71. Drakatos, P. et al. First rapid eye movement sleep periods and sleep-onset rapid eye movement periods in sleep-stage sequencing of hypersomnias. *Sleep Med.* **14**, 897–901 (2013).
72. Liu, Y. et al. Altered sleep stage transitions of REM sleep: a novel and stable biomarker of narcolepsy. *J. Clin. Sleep Med.* **11**, 885–894 (2015).
73. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* (The MIT Press, Cambridge, 2006).
74. Hensman, J., Matthews, A. & Ghahramani, Z. Scalable variational Gaussian process classification. In *18th International Conference on Artificial Intelligence and Statistics (AISTATS)* (PMLR, San Diego, CA, 2015).
75. Matthews, A. G. D. G., Nickson, T., Boukouvalas, A. & Hensman, J. GPflow: a Gaussian Process Library using TensorFlow. *J. Mach. Learn. Res.* **18**, 1–6 (2017).

Acknowledgements

This research was mostly supported by a grant from Jazz Pharmaceuticals to E.M. Additional funding came from: NIH grant R01HL62252 (to P.E.P.); Ministry of Science and Technology 2015CB856405 and National Foundation of Science of China 81420108002,81670087 (to F.H.); H. Lundbeck A/S, Lundbeck Foundation, Technical University of Denmark and Center for Healthy Aging, University of Copenhagen (to P.J. and H.B.D.S). Additional support was provided by the Klarman Family, Otto Mønsted, Stibo, Vera & Carl Johan Michaelsens, Knud Højgaards, Reinholdt W. Jorck and Hustrus and Augustinus Foundations (to A.N.O.). We thank Briana Gomez for helping on performing the last corrections on this manuscript.

Author contributions

J.B.S. and A.N.O. participated in the design of the study, conducted most of the analyses and did most of the publication writing. M.O., A.A., E.B.L., H.E.M., O.C., D.P. and L.L. assisted in many of the analyses and data organization, plus helped edit the manuscript. F.H., H.Y., Y.L.S., Y.D., S.S., L.B., B.H., A.S., S.C.H., T.W.K., F.P., G.P., S.V., E.A., S.T.K., P.K.S., C.K. and P.E.P. contributed essential datasets, helped organization and gave feedback to manuscript content. P.J. and H.B.D.S. participated in the design of the study, supervised analyses and assisted publication writing. P.J. furthermore contributed datasets to the study. E.M. contributed datasets, participated in the design of the study, supervised the analyses and did most of the publication writing.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-07229-3.

Competing interests: E.M. has received Jazz Pharmaceuticals contract, clinical trial and gift funding as principal investigator at Stanford University. He also consulted for

Idorsia and Merck, consulted or presented clinical trial results for Jazz Pharmaceuticals at congresses, this resulting in trip reimbursements and honoraria never exceeding 5000 dollars per year. G.P. has been on advisory boards for UCB, Jazz, Bioprojet and Idorsia. F.P. received a fee from UCB for speaking at a symposium, and a congress subscription from Bioprojet. The remaining authors declare no competing interests.

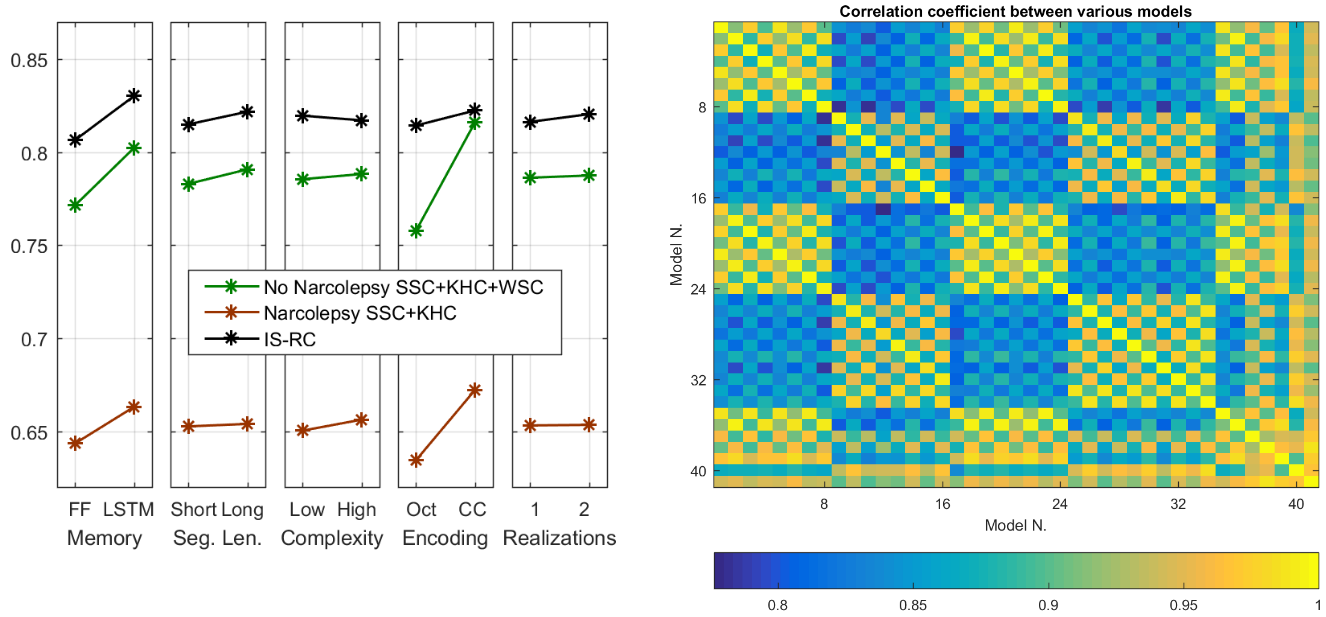
Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

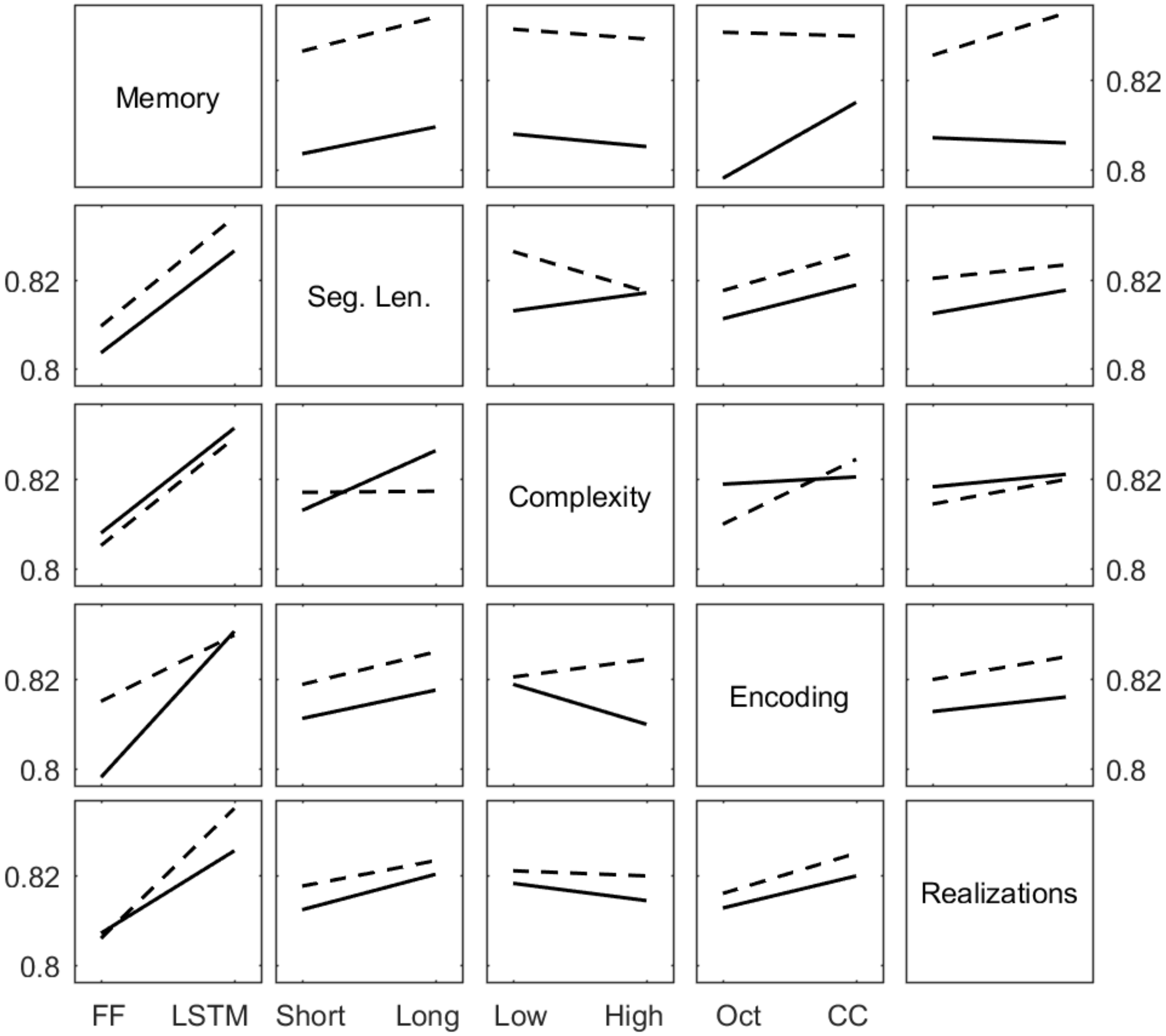
Supplementary Material to Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy

Stephansen and Olesen et al.

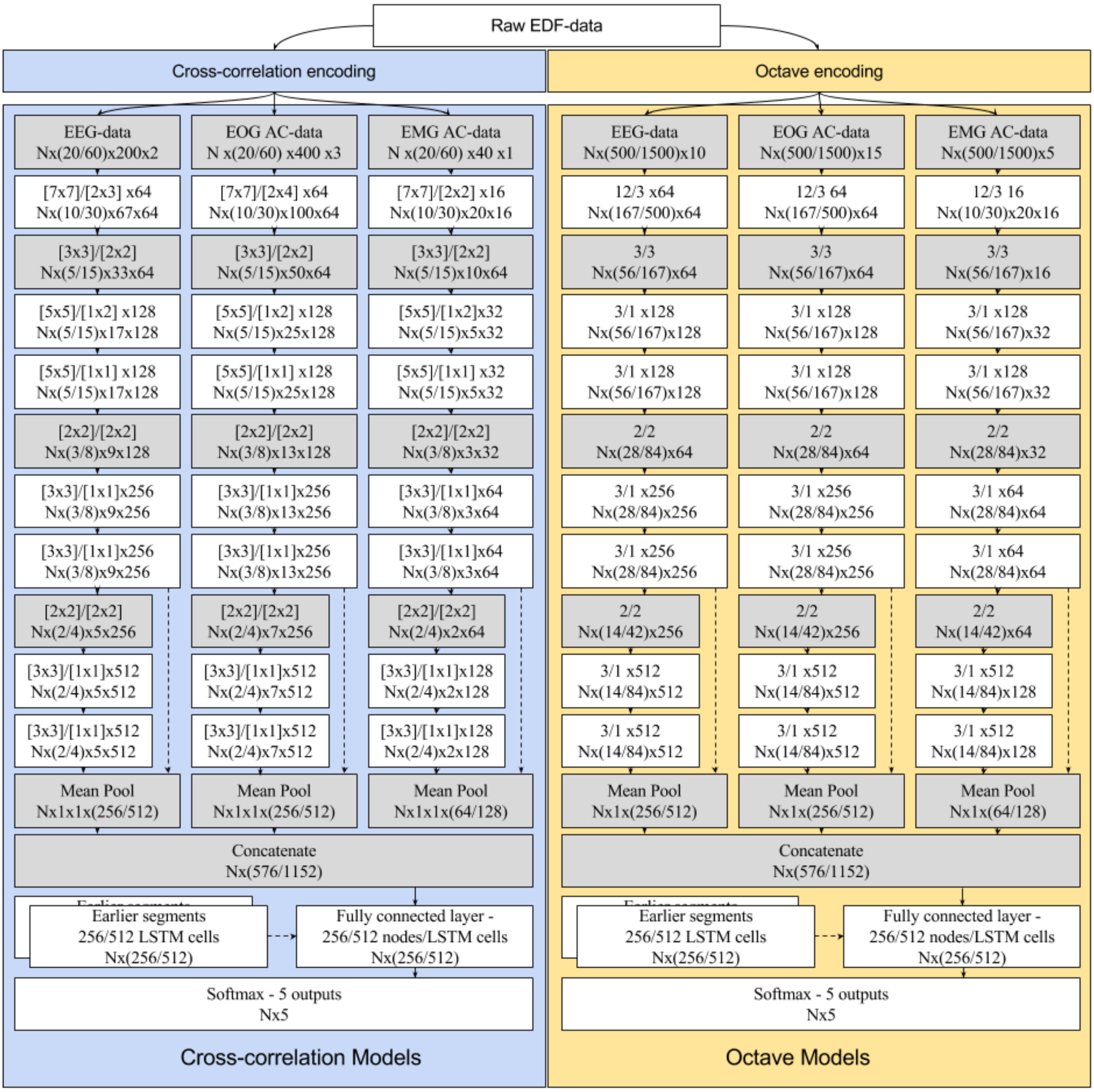
SUPPLEMENTARY FIGURES



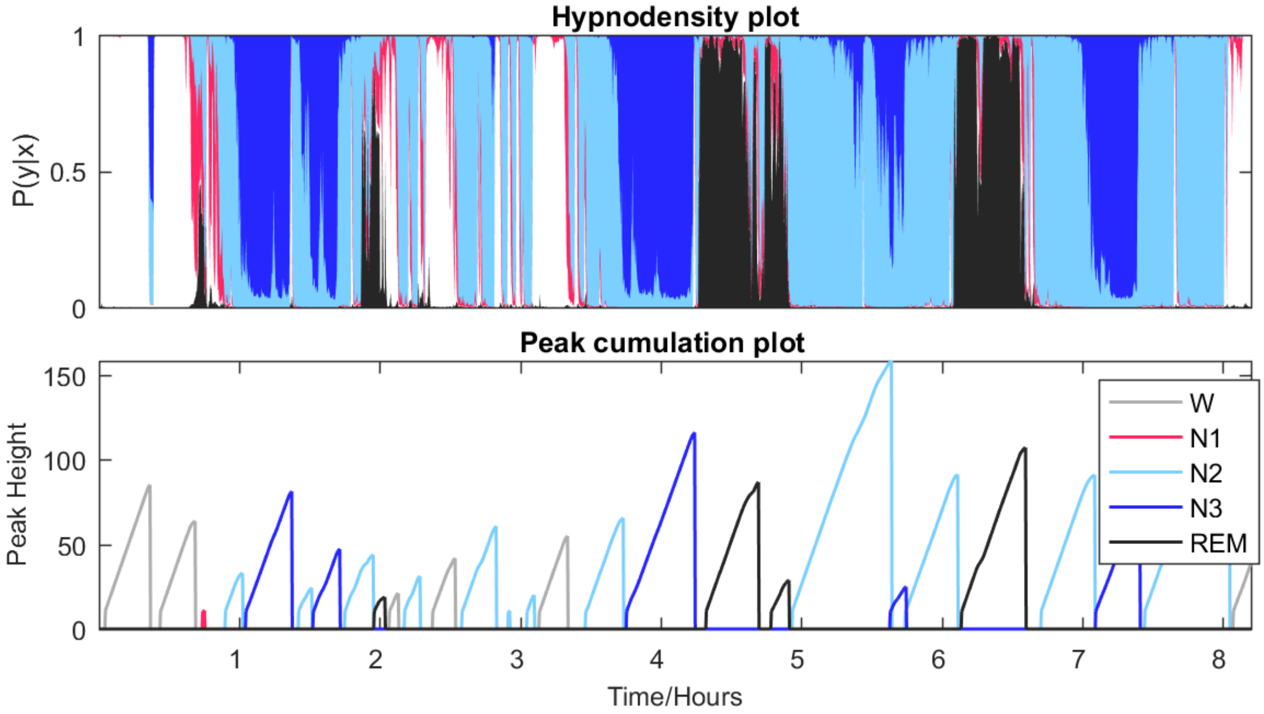
Supplementary Figure 1: Comparisons of machine learning models. Left: Comparisons of the effect on accuracy by each factor at different settings on IS-RC data, SSC and KHC narcolepsy subjects, and the remaining SSC, KHC and WSC subjects used for testing. Right: Correlation matrix showing similarities in different model predictions, where 0 means signals are independent, and 1 means signals are completely correlated. Models number (N) 1-32 are single models, and 33-41 are ensembles. The models vary on 5 parameters, each at two levels, in the following order: Memory – FF or LSTM (1), segment length (Seg. Len.) – 5 s or 15 s (2), complexity – high or low (3), encoding – CC or octave (4), realizations – 1 or 2 (5). Ensembles are as described in Supplementary Table 8: All FF octave models (33), all LSTM octave models (34), all FF CC models (35), all LSTM CC models (36), all FF models (37), all LSTM models (38), all CC models (39), all octave models (40), all models (41).



<u>Supplementary Figure 2</u>: Interaction of different factors and their dependence on accuracy. The IS-RC data was used for this analysis. The solid and dashed lines indicate factors along the rows on levels 1 and 2, respectively.

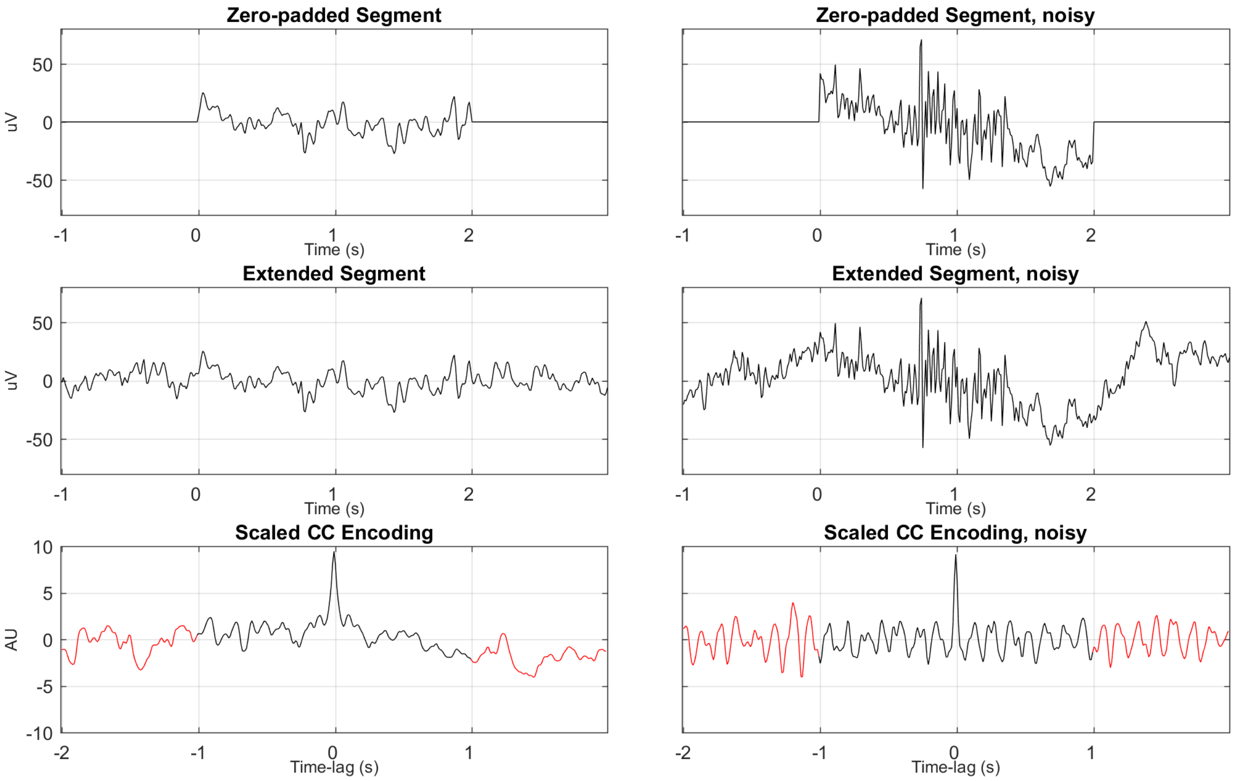

Raw EDF-data
Cross-correlation encoding
Octave encoding
EEG-data Nx(20/60)x200x2
EOG AC-data N x(20/60) x400 x3
EMG AC-data N x(20/60) x40 x1
[7x7]/[2x3] x64 Nx(10/30)x67x64
[7x7]/[2x4] x64 Nx(10/30)x100x64
[7x7]/[2x2] x16 Nx(10/30)x20x16
[3x3]/[2x2] Nx(5/15)x33x64
[3x3]/[2x2] Nx(5/15)x50x64
[3x3]/[2x2] Nx(5/15)x10x64
[5x5]/[1x2] x128 Nx(5/15)x17x128
[5x5]/[1x2] x128 Nx(5/15)x25x128
[5x5]/[1x2]x32 Nx(5/15)x5x32
[5x5]/[1x1] x128 Nx(5/15)x17x128
[5x5]/[1x1] x128 Nx(5/15)x25x128
[5x5]/[1x1] x32 Nx(5/15)x5x32
[2x2]/[2x2] Nx(3/8)x9x128
[2x2]/[2x2] Nx(3/8)x13x128
[2x2]/[2x2] Nx(3/8)x3x32
[3x3]/[1x1]x256 Nx(3/8)x9x256
[3x3]/[1x1]x256 Nx(3/8)x13x256
[3x3]/[1x1]x64 Nx(3/8)x3x64
[3x3]/[1x1]x256 Nx(3/8)x9x256
[3x3]/[1x1]x256 Nx(3/8)x13x256
[3x3]/[1x1]x64 Nx(3/8)x3x64
[2x2]/[2x2] Nx(2/4)x5x256
[2x2]/[2x2] Nx(2/4)x7x256
[2x2]/[2x2] Nx(2/4)x2x64
[3x3]/[1x1]x512 Nx(2/4)x5x512
[3x3]/[1x1]x512 Nx(2/4)x7x512
[3x3]/[1x1]x128 Nx(2/4)x2x128
[3x3]/[1x1]x512 Nx(2/4)x5x512
[3x3]/[1x1]x512 Nx(2/4)x7x512
[3x3]/[1x1]x128 Nx(2/4)x2x128
Mean Pool Nx1x1x(256/512)
Mean Pool Nx1x1x(256/512)
Mean Pool Nx1x1x(64/128)
Concatenate Nx(576/1152)
Earlier segments
Earlier segments 256/512 LSTM cells Nx(256/512)
Fully connected layer - 256/512 nodes/LSTM cells Nx(256/512)
Softmax - 5 outputs Nx5
Cross-correlation Models
EEG-data Nx(500/1500)x10
EOG AC-data Nx(500/1500)x15
EMG AC-data Nx(500/1500)x5
12/3 x64 Nx(167/500)x64
12/3 64 Nx(167/500)x64
12/3 16 Nx(10/30)x20x16
3/3 Nx(56/167)x64
3/3 Nx(56/167)x64
3/3 Nx(56/167)x16
3/1 x128 Nx(56/167)x128
3/1 x128 Nx(56/167)x128
3/1 x128 Nx(56/167)x32
3/1 x128 Nx(56/167)x128
3/1 x128 Nx(56/167)x128
3/1 x128 Nx(56/167)x32
2/2 Nx(28/84)x64
2/2 Nx(28/84)x64
2/2 Nx(28/84)x32
3/1 x256 Nx(28/84)x256
3/1 x256 Nx(28/84)x256
3/1 x64 Nx(28/84)x64
3/1 x256 Nx(28/84)x256
3/1 x256 Nx(28/84)x256
3/1 x64 Nx(28/84)x64
2/2 Nx(14/42)x256
2/2 Nx(14/42)x256
2/2 Nx(14/42)x64
3/1 x512 Nx(14/84)x512
3/1 x512 Nx(14/84)x512
3/1 x512 Nx(14/84)x128
3/1 x512 Nx(14/84)x512
3/1 x512 Nx(14/84)x512
3/1 x512 Nx(14/84)x128
Mean Pool Nx1x(256/512)
Mean Pool Nx1x(256/512)
Mean Pool Nx1x(64/128)
Concatenate Nx(576/1152)
Earlier segments
Earlier segments 256/512 LSTM cells Nx(256/512)
Fully connected layer - 256/512 nodes/LSTM cells Nx(256/512)
Softmax - 5 outputs Nx5
Octave Models

Supplementary Figure 3: Specifications of each network configuration. Each block represents an operation; with white blocks require multiplications and adding, whereas grey blocks are pooling or concatenations, default being max pooling. The top row of each block describes the size of the window and its stride, and the bottom row describes the size of the output. In this output, N is the length of a sequence, the second dimension is the segment length, and if a fourth dimension is present (CC models), the third dimension originally represents the size of the correlation function. The last dimension is the number of features in that layer. Models with a low complexity skip the third max pooling block, and go straight to mean pooling.



.

Supplementary Figure 4: Peak cumulation plot. It visualizes how hypnodensity-derived features are calculated (See Supplementary Table 10). Color codes: White – wake, red – N1, light blue – N2, dark blue – N3, black – REM



<u>Supplementary Figure 5:</u> Implementation of CC encoding. CC encoding of a noisy (right) and less noisy (left) signal. The central part of the encoding, representing areas of full overlap between correlated signals, is kept; the red part is discarded.

SUPPLEMENTARY TABLES

Supplementary Table 1: Description of the various cohorts included in this study and how they were used.

Cohort	Age (μ ± σ)	BMI (μ ± σ)	Sex (% male)	Sleep scoring		Narcolepsy biomarker			% narco	% hypersomnia	Use
				Train	Test	Train	Test	Replication			
WSC	59.7 ± 8.4	31.6 ± 7.1	53.1	1,086 (2,167 PSGs)	286	170	116	None	0.0	0.0	Training and testing of sleep scoring models and narcolepsy biomarker.
SSC	45.4 ± 13.8	23.9 ± 6.5	59.4	617	277	139	112	None	11.6	1.8	Training and testing of sleep scoring models and narcolepsy biomarker.
KHC	29.1 ± 13.2	24.1 ± 4.3	58.6	None	160	87	71	None	45.8	54.2	Sleep scoring testing, and training and testing of narcolepsy biomarker.
AHC	34.5 ± 13.8	25.9 ± 4.9	54.0	None	None	42 (76 PSGs)	44 (84 PSGs)	None	52.3	47.7	Training and testing of narcolepsy biomarker. 86 subjects had the first PSG recorded, and 75 had an additional second PSG. A subject was used for either training or testing.
IS-RC	51.1 ± 4.2	32.9 ± 9.2	0.0	None	70	None	None	None	0.0	0.0	Scored by 6 different scorers. Final assessment and validation of predictive performance for sleep scoring.
JCTS	53.2 ± 9.8	31.0 ± 4.4	57.1	None	None	7	None	None	100.0	0.0	Training of narcolepsy biomarker.
IHC	33.7 ± 17.6	-	56.7	None	None	87	61	None	47.3	50.0	Training and testing of narcolepsy biomarker.
DHC	33.4 ± 14.8	24.8 ± 4.9	50.0	None	None	79	None	None	26.6	48.1	Training of narcolepsy biomarker.
FHC	28.8 ± 15.2	24.4 ± 8.1	59.0	None	None	None	None	122	51.6	18.0	Replication of narcolepsy biomarker in never seen datasets
CNC	28.5 ± 16.9	23.2 ± 11.5	51.3	None	None	None	None	199	34.2	0.0	Replication of narcolepsy biomarker in never seen datasets
Total subjects				1,703	793	611	404	321			
Total PSGs				2,784	793	645	444	321			

% narco. = % of cohort with type 1 narcolepsy; % hypersomnia= % with idiopathic hypersomnia or narcolepsy type 2 (high pretest probability cohort)

Supplementary Table 2: A cumulative assessment of the scorers.

		Consensus					
	Stages	**Wake**	**N1**	**N2**	**N3**	**REM**	
Accumulation of Individual Scorers	Wake	13.28% 13.25%	1.04% 0.98%	0.86% 0.87%	0.08% 0.08%	0.23% 0.22%	0.86 0.86
	N1	0.79% 0.88%	3.36% 3.61%	1.23% 1.42%	0.03% 0.03%	0.29% 0.31%	0.59 0.58
	N2	0.87% 0.84%	2.46% 2.30%	44.66% 45.48%	4.89% 5.92%	0.85% 0.84%	0.83 0.82
	N3	0.05% 0.05%	0.02% 0.02%	2.58% 1.54%	6.45% 5.41%	0.002% 0.002%	0.71 0.77
	REM	0.32% 0.31%	1.00% 0.97%	1.14% 1.16%	0.03% 0.04%	13.46% 13.46%	0.84 0.84
		0.87 0.86	0.43 0.46	0.88 0.90	0.56 0.47	0.91 0.91	**0.81** **0.81**

The top row in every cell displays the un-weighed consensus, and the bottom row displays the weighed consensus. The values in the diagonal indicate a match between scorer and consensus. The total number of scored epochs were 324,978

Supplementary Table 3: Average relative model variance, standardized to a correct wakefulness prediction, when compared to the scoring consensus. On average, the sleep classification model shows lower variance in the diagonal, which translates to a higher certainty on predicted true positives.

		Model Predictions				
	Stages	**Wake**	**N1**	**N2**	**N3**	**REM**
Consensus	Wake	1.00	1.16	2.25	2.12*	3.74
	N1	1.58	0.89	1.08	0.03*	1.29
	N2	3.80	1.33	0.51	0.99	1.45
	N3	0.92*	NaN*	1.36	0.58	NaN*
	REM	3.58	1.89	1.93	NaN*	1.06

*Fewer than five observations.

<u>Supplementary Table 4</u>: ANOVA comparing accuracy for subjects with and without various sleep disorders.

Condition	Source	Sum of squares	Degrees of freedom	p-value	Delta Mean accuracy	
Insomnia (N = 333) N_{Insomnia} = 134	Cohort	0.30	2	$3.69\cdot10^{-21}$	Present	0.04
	Age	0.0026	2	0.62		
	Sex	0.0060	1	0.139		
	Condition	0.0003	1	0.75		
	Error	0.89	326			
OSA (N = 683)	Cohort	2.85	2	$2.81\cdot10^{-82}$		
N_{None} = 297	Age	0.045	2	0.020	None	-
N_{Mild} = 167	Sex	0.0018	1	0.57	Mild	0.04
N_{Moderate} = 118	Condition	0.097	3	$7.53\cdot10^{-4}$	Moderate	0.03
N_{Severe} = 101	Error	3.82	674		Severe	0.00
RLS (N = 580) N_{RLS} = 136	Cohort	2.16	2	$6.50\cdot10^{-54}$	Present	0.08
	Age	0.056	2	0.020		
	Sex	0.011	1	0.22		
	Condition	0.016	1	0.13		
	Error	4.05	573			
PLMI (N = 288)	Cohort	-	-	-		
N_{None} = 120	Age	0.0027	1	0.31	None	-
N_{Mild} = 80	Sex	0.0014	1	0.45	Mild	0.00
N_{Moderate} = 55	Condition	0.011	3	0.22	Moderate	-0.01
N_{Severe} = 33	Error	3.9297	282		Severe	-0.02
Narcolepsy (N = 729) $N_{\text{Narcolepsy}}$ = 98	Cohort	2.05	2	$1.63\cdot10^{-65}$	Present	-0.15
	Age	0.13	2	$6.73\cdot10^{-6}$		
	Sex	0.018	1	0.070		
	Condition	0.368	1	$1.77\cdot10^{-15}$		
	Error	4.01	722			
Overall (N = 729)	Cohort	2.97	2	$6.10\cdot10^{-82}$		
	Age	0.065	2	0.0047		
	Sex	0.010	1	0.19		
	Error	4.38	723			

The model used is the ensemble of all CC models. Each analysis is done separately to account for missing values. Cohorts are the SSC, WSC, KHC and AHC. Age is grouped as age<30, 30≤age<50 and age≥50. OSA is grouped as AHI<5, 5≤AHI<15, 15≤AHI<30 and AHI≥30. PLM is grouped as PLMI <5, 5≤ PLMI <15, 15≤ PLMI <30 and PLMI ≥30.

Supplementary Table 5: Selection frequency and descriptions of each of the 38 features included in the Gaussian process model used for narcolepsy prediction.

#	Feature # in supplementary Table 10.	Stage Combination	Relative selection frequency
1	12	W, N2, REM	1
2	Nightly SOREMPs (REM latency ≤ 15 min)		0.91
3	15	W	0.82
4	6	REM	0.82
5	2	W	0.68
6	2	N2, REM	0.68
7	14	W, N2	0.68
8	13	W, N1	0.64
9	5	N3	0.59
10	5	REM	0.59
11	13	N1, N2	0.59
12	8	N1	0.55
13	11	N1	0.55
14	7	W, N1, REM	0.55
15	5	W, N1, N3	0.55
16	6	W, N1, N3	0.55
17	1	W, N1, N2, REM	0.55
18	Hypnodensity sleep stage bout transitions: N2 to N3		0.55
19	Accumulation of the wakeful periods ≤ 15 minutes		0.50
20	Hypnodensity sleep stage bout transitions: W/N1 to REM		0.50
21	11	N3, REM	0.45
22	2	N1, REM	0.45
23	7	W, N2, N3	0.45
24	12	W	0.41
25	2	N1	0.41
26	12	N2	0.41
27	14	N2	0.41
28	7	N2, REM	0.41
29	8	N2, REM	0.41
30	6	N1, N2	0.41
31	15	N1, N2	0.41
32	15	W, N3	0.41
33	12	W, N1	0.41
34	5	W, N2, REM	0.41
35	1	W, N1, N3, REM	0.41
36	1	W, N1, N2, N3, REM	0.41
37	Accumulation of REM epochs following wakeful periods		0.41
38	Hypnodensity sleep stage bout transitions: N2 to REM		0.41

Supplementary Table 6: Descriptive statistics on the evaluation of the narcolepsy biomarker in models with and without the HLA biomarker. Performance on models with HLA typing is reported for regular threshold and optimized threshold, since the ROC curve is changed dramatically by adding HLA. Mean value and 95% confidence interval. PPV and NPV are positive and negative predictive value, respectively.

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Number of PSGs	T1N fraction
Test (T)	0.95 0.92-0.97	0.91 0.84-0.96	0.96 0.93-0.98	0.88 0.80-0.93	0.97 0.95-0.99	444	0.24
Replication (R)	0.92 0.88-0.95	0.93 0.87-0.97	0.91 0.87-0.95	0.87 0.80-0.93	0.95 0.92-0.98	321	0.28
T+R, HLA	0.96 0.94-0.97	0.90 0.84-0.93	0.99 0.98-1.00	0.97 0.94-0.99	0.95 0.93-0.97	584	0.31
T+R, HLA, optimized	0.94 0.92-0.96	0.94 0.90-0.97	0.94 0.92-0.96	0.88 0.83-0.92	0.97 0.95-0.99	584	0.31
High pre-test (HPT), no HLA.	0.91 0.87-0.94	0.90 0.86-0.94	0.92 0.86-0.96	0.94 0.91-0.97	0.86 0.80-0.91	335	0.61
HPT, HLA	0.93 0.90-0.95	0.90 0.84-0.93	0.98 0.96-1.00	0.99 0.97-1.00	0.85 0.79-0.91	296	0.61
HPT, HLA, optimized	0.93 0.90-0.95	0.94 0.90-0.97	0.90 0.85-0.95	0.94 0.90-0.97	0.90 0.85-0.95	296	0.61

Supplementary Table 7: Confusion matrix on the SSC and KHC data, displaying the relationship between scorer and the ensemble estimate as a fraction of the amount of data in total.

		Target					
	Stages	**Wake**	**N1**	**N2**	**N3**	**REM**	
Model prediction	Wake	13.94% 8.02% $2.08\cdot10^{-5}$	0.40% 0.54% 0.085	1.46% 1.59% 0.54	0.04% 0.07% 0.01	0.43% 0.59% 0.097	0.86 0.74
	N1	2.58% 3.59% 0.014	1.51% 1.53% 0.916	3.64% 2.70% 0.024	0.08% 0.13% 0.095	1.14% 1.57% 0.011	0.17 0.16
	N2	2.18% 4.07% $4.59\cdot10^{-6}$	1.30% 2.79% $4.86\cdot10^{-12}$	42.55% 38.59% 0.002	2.06% 1.94% 0.714	1.73% 2.18% 0.090	0.85 0.78
	N3	0.02% 0.02% 0.872	0.002% 0.003% 0.582	2.68% 4.05% 0.001	5.84% 7.67% 0.023	0.004% 0.009% 0.357	0.68 0.65
	REM	0.99% 3.03% $4.07\cdot10^{-12}$	0.36% 0.71% $1.43\cdot10^{-5}$	1.81% 1.91% 0.674	0.05% 0.06% 0.753	13.01% 12.64% 0.588	0.80 0.69
		0.71 0.43	0.42 0.27	0.82 0.79	0.72 0.78	0.80 0.74	0.77 0.68

The top row of each cell is data from non-narcoleptics, the second row is from narcoleptics, and the bottom row is the p-value, indicating whether there is a significant difference in the two means. 98 narcolepsy subjects and 500 non-narcolepsy subjects were used for the analysis.

Supplementary Table 8: Models tested. 32 single models are tested, and 9 ensembles, totaling 41 models.

			Single models		
	Memory	**Seg. Len.**	**Complexity**	**Encoding**	**Realizations**
Configuration 1	Simple FF	5 s	Low	Octave	1
Configuration 2	LSTM	15 s	High	CC	2

Ensembles

Parameters included	All Oct FF	All Oct LSTM	All CC FF	All CC LSTM	All FF	All LSTM	All Oct models	All CC models	All models
N. models	8	8	8	8	16	16	16	16	32

Supplementary Table 9: The number of stage combinations, and the number of features this leads to.

	Single stage	Two stages	Three stages	Four stages	Five stages	Additional	Total
Combinations	5	10	10	5	1		31
Features	75	150	150	75	15	16	481

Supplementary Table 10: Description of each feature, how it is calculated, and how it is numerated.

#	Description of what is expressed	Formula
1	General prevalence of a value	$\log\left(\frac{1}{N}\sum_{seg=1}^{N}\mathbf{\Phi}(\mathcal{C}_k)\right)$
2	Highest achieved value, measured as the distance from the highest value possible.	$-\log\left(1-\mathrm{maximum}\left(\mathbf{\Phi}(\mathcal{C}_k)\right)\right)$
3	Measures average fluctuations in value.	$\log\left(\frac{1}{N}\sum_{seg=1}^{N}\left\lvert\frac{d\mathbf{\Phi}(\mathcal{C}_k)}{dseg}\right\rvert\right)$
4	Log of Shannon entropy, calculated through a wavelet decomposition, where s_i contains the wavelet decompositions of $\mathbf{\Phi}(\mathcal{C}_k)$. Measures the amount of information contained in the signal, i.e. how many different values are achieved.	$log\left(\frac{-\sum_i s_i^2\log s_i^2}{N}\right)$
5 **6** **7** **8**	Time until 5%, 10%, 30% or 50% of the maximum value has been achieved.	$\log\left(\mathrm{first}_{arg>5\%,10\%,30\%,50\%}\left(\frac{\mathrm{cumsum}(\mathbf{\Phi}(\mathcal{C}_k))}{\mathrm{sum}\left(\mathbf{\Phi}(\mathcal{C}_k)\right)}\right)\cdot 30\right)$
9	Maximum value achieved weighed by the mean prevalence.	$\sqrt{\left(\mathrm{maximum}\left(\mathbf{\Phi}(\mathcal{C}_k)\right)\cdot\mathrm{mean}(\mathbf{\Phi}(\mathcal{C}_k))\right)}$
10	Average fluctuations of value weighed by mean prevalence.	$\left(\frac{1}{N}\sum_{seg=1}^{N}\left\lvert\frac{d\mathbf{\Phi}(\mathcal{C}_k)}{dseg}\right\rvert\right)\cdot\mathrm{mean}(\mathbf{\Phi}(\mathcal{C}_k))$
11	Shannon entropy weighed by mean prevalence.	$log\left(\frac{-\sum_i s_i^2\log s_i^2}{N}\cdot\mathrm{mean}(\mathbf{\Phi}(\mathcal{C}_k))\right)$
12 **13** **14** **15**	Time until 5%, 10%, 30% or 50% of the maximum value has been achieved weighed by mean prevalence.	$\sqrt{\left(\mathrm{first}_{arg>5\%,10\%,30\%,50\%}\left(\frac{\mathrm{cumsum}(\mathbf{\Phi}(\mathcal{C}_k))}{\mathrm{sum}\left(\mathbf{\Phi}(\mathcal{C}_k)\right)}\right)\cdot 30\mathrm{mean}(\mathbf{\Phi}(\mathcal{C}_k))\right)}$

Each individual feature is scaled by subtracting the mode dividing by the difference between the 85th and 15th percentile. Each value was assessed visually to ensure that the transformations and scaling was done optimally. cumsum is the culminative sum.

D

PAPER IV

TITLE: Towards a Flexible Deep Learning Method for Automatic Detection of Clinically Relevant Multi-Modal Events in the Polysomnogram

AUTHORS: Alexander Neergaard Olesen, Stanislas Chambon, Valentin Thorey, Poul Jennum, Emmanuel Mignot, and Helge B. D. Sorensen

CONFERENCE: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)

STATUS: Published

DOI: 10.1109/EMBC.2019.8856570

FULL CITATION: A. N. Olesen, S. Chambon, V. Thorey, P. Jennum, E. Mignot and H. B. D. Sorensen, "Towards a Flexible Deep Learning Method for Automatic Detection of Clinically Relevant Multi-Modal Events in the Polysomnogram," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, 2019, pp. 556-561.

Towards a Flexible Deep Learning Method for Automatic Detection of Clinically Relevant Multi-Modal Events in the Polysomnogram

Alexander Neergaard Olesen[†,1,2,3], *Member, IEEE*, Stanislas Chambon[4,5], Valentin Thorey[5], Poul Jennum[3], Emmanuel Mignot[2] and Helge B. D. Sorensen[3], *Senior Member, IEEE*

Abstract— **Much attention has been given to automatic sleep staging algorithms in past years, but the detection of discrete events in sleep studies is also crucial for precise characterization of sleep patterns and possible diagnosis of sleep disorders. We propose here a deep learning model for automatic detection and annotation of arousals and leg movements. Both of these are commonly seen during normal sleep, while an excessive amount of either is linked to disrupted sleep patterns, excessive daytime sleepiness impacting quality of life, and various sleep disorders. Our model was trained on 1,485 subjects and tested on 1,000 separate recordings of sleep. We tested two different experimental setups and found optimal arousal detection was attained by including a recurrent neural network module in our default model with a dynamic default event window (F1 = 0.75), while optimal leg movement detection was attained using a static event window (F1 = 0.65). Our work show promise while still allowing for improvements. Specifically, future research will explore the proposed model as a general-purpose sleep analysis model.**

I. INTRODUCTION

Analysis of sleep patterns is performed manually by experts in sleep clinics using rules and guidelines defined by the American Academy of Sleep Medicine recently updated in 2018 [1]. These guidelines outline technical and clinical best practices when performing routine polysomnography (PSG), which is an overnight recording of electroencephalography (EEG), electrooculography (EOG), electromyography (EMG) electrocardiography (ECG), respiratory effort and peripheral limb activity. Expert technicians and somnologists use these physiological variables to analyse sleep patterns and diagnose sleep disorders based on key metrics and indices, such as total sleep time, amount of sleep spent in various sleep stages, and the observed number of discrete events per hour of sleep. Specifically, the number of arousals (short awakenings during sleep, $<$15 s), non-periodic and periodic leg movements (PLM), and the number of apnea events per hour of sleep are summarized in the arousal index (AI), periodic leg movements index (PLMI) and apnea/hypopnea index (AHI), the latter of which is a combination of apneic (no/obstructed respiratory effort) and hypopneic (reduced respiratory effort) events. Excessive amounts of these events are disruptive to normal sleep, which can lead to patient complaints of excessive daytime sleepiness [2], which in turn is linked to an increase in e.g. automotive accidents and reduced quality of life [3]. Increased number of PLMs is also linked to other sleep disorders such as restless legs syndrome, and periodic leg movement disorder [4], [5].

Correct diagnosis of sleep disorders is predicated on precise scoring of sleep stages as well as accurate scoring of these discrete sleep events. However, the current gold standard of manual analysis by experienced technicians is inherently biased and inconsistent.Several studies have shown low inter-rater reliability on both the scoring of sleep stages [6]–[8], arousals [9], and respiratory events [10]. Furthermore, manual analysis of PSGs is time-consuming and prone to scorer fatigue. Thus, there is a need for efficient systems that provide deterministic and reliable scorings of sleep studies.

Several recent studies have already explored automatic classification of sleep stages in large cohorts with good results [11]–[15], however, the reliable and consistent detection and classification of discrete PSG events in large cohorts remain largely unexplored.

Recent studies on certain microevents in sleep have indicated that sleep spindles and K-complexes can be reliably detected and annotated with start time and duration using deep learning methods [16], [17]. Specifically, these studies proposed a single-shot event detection algorithm, that parallels the YOLO and SSD algorithms used for object detection in 2D images [18], [19], however, they were limited in scope by detecting events only at the EEG level, and did not explicitly take advantage of the temporal connection of the detected events. Additionally, experiments were carried out on a small-scale database [16].

In this study, we focused on the detection of arousals (AR) and leg movements (LM). These events arise from highly distinct physiological sources, EEG and leg EMG, while ARs are also visible in the EOG and chin EMG. These events are important for the precise characterization of sleep patterns and possible diagnosis of sleep disorders, and an accurate detection is therefore of high interest. We extend previous work in [16], [17] by 1) preprocessing and analysing multiple input signals at the same time, and 2) taking into account important temporal context using recurrent neural networks.

Research supported by the Klarman Family Foundation, Technical University of Denmark, and University of Copenhagen with supporting grants from Reinholdt W. Jorck & Wife's Foundation, Knud Højgaard Foundation, Otto Mønsted Foundation, Vera & Carl Michaelsens Foundation, Augustinus Foundation, and Stibo Foundation.

[†] Corresponding author: alexno@stanford / aneol@dtu.dk

[1] Department of Health Technology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark.

[2] Center for Sleep Sciences and Medicine, Stanford University, Palo Alto, CA 94304, USA.

[3] Danish Center for Sleep Medicine, University Hospital Copenhagen, 2600 Glostrup, Denmark

[4] LTCI Telecom ParisTech, Universite Paris-Saclay, Paris, France.

[5] Research & Algorithms Team, Dreem, Paris, France.

978-1-5386-1311-5/19/$31.00 ©2019 IEEE

TABLE I: MrOS subset demographics. Significant p-values at $\alpha = 0.05$ are shown in bold.

	TRAIN	EVAL	TEST	p-value
N	1,485	165	1000	-
Age (years)	76.4 ± 5.5	76.6 ± 4.9	76.4 ± 5.6	0.631
BMI ($\mathrm{kg\,s^{-2}}$)	27.2 ± 3.8	27.2 ± 3.4	27.1 ± 3.7	0.879
AHI ($\mathrm{h^{-1}}$)	12.8 ± 12.9	10.6 ± 11.8	11.9 ± 12.8	**0.029**
AI ($\mathrm{h^{-1}}$)	23.6 ± 11.5	24.1 ± 12.2	23.4 ± 11.8	0.607
PLMI ($\mathrm{h^{-1}}$)	34.8 ± 37.0	37.8 ± 38.9	37.3 ± 38.0	0.204

Furthermore, we apply our model on a larger database than previous studies.

II. DATA

A. MrOS Sleep Study

The MrOS Sleep Study is a part of the larger Osteoporotic Fractures in Men Study with the objective of researching the links between sleep disorders, fractures, cardiovascular disease and mortality in older males (> 65 years) [20]–[22]. Between 2003 and 2005, 3,135 of the original 5,994 participants were recruited to undergo full-night PSG recording at six centers in the US at two separate visits (visit 1 and visit 2) with following 3 to 5-day actigraphy studies at home. The resulting PSG studies were subsequently scored by experienced sleep technicians for standard sleep variables including sleep stages, leg movements, arousals, and respiratory events.

B. Included events and signals

In this study, we only considered the detection of two PSG events, arousals and leg movements. These events are characterized by a start time and a duration, which we extracted from 2,907 PSG studies from visit 1 available from the National Sleep Research Resource repository [23], [24]. From each PSG study, we extracted left and right central EEG, left and right EOG, chin EMG, and EMG from the left and right anterior tibialis. EEG and EOG channels were referenced to the contralateral mastoid process, while a leg EMG channel was synthesized by referencing left to right. Any PSG without the full set of channels or without any event scoring was eliminated from further analysis.

C. Subset demographics and partitioning

In total, 2,650 out of the 2,907 PSGs available from visit 1 were included in this study. These were partitioned into TRAIN, EVAL, and TEST sets of sizes 1,485, 165, and 1,000 studies, respectively. A subset of key demographic and PSG variables are presented in Table I.

III. METHODS

A. Signal preprocessing

All signals were resampled to $f_s = 128\,\mathrm{Hz}$ using polyphase filtering with a Kaiser window ($\beta = 5.0$) before subsequent filtering according to AASM criteria. Briefly, EEG and EOG channels were subjected to a 4th order Butterworth band pass filter with cutoff frequencies $[0.3, 35.0]$ Hz, while chin and leg EMG channels were filtered with a 4th order Butterworth high pass filter with a 10 Hz cutoff frequency. All filters employed zero-phase filtering. Lastly, each channel was normalized by subtracting the channel mean and dividing by the channel standard deviation across the entire night.

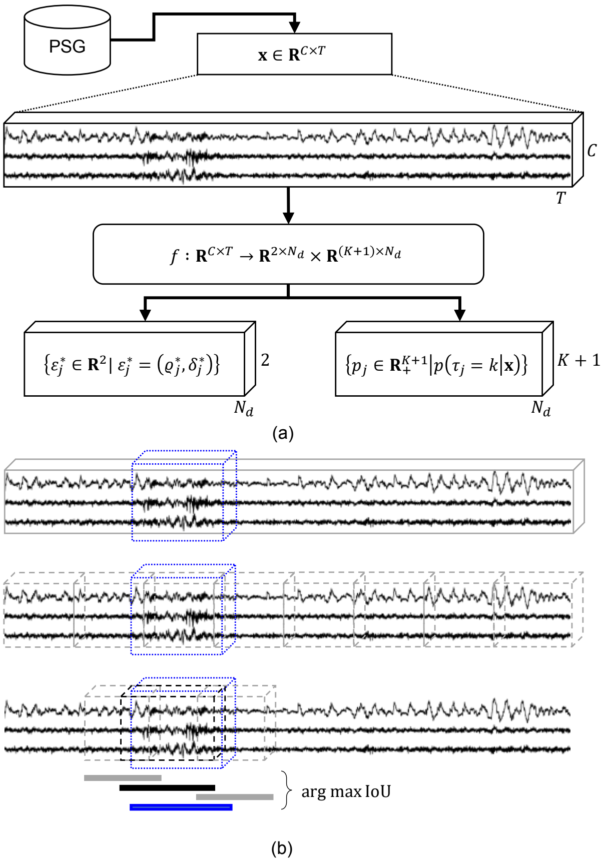

PSG
$\mathbf{x} \in \mathbf{R}^{C \times T}$
C
T
$f : \mathbf{R}^{C \times T} \to \mathbf{R}^{2 \times N_d} \times \mathbf{R}^{(K+1) \times N_d}$
$\{\varepsilon_j^* \in \mathbf{R}^2 \mid \varepsilon_j^* = (\varrho_j^*, \delta_j^*)\}$
2
N_d
$\{p_j \in \mathbf{R}_+^{K+1} \mid p(\tau_j = k \mid \mathbf{x})\}$
$K+1$
N_d
(a)
arg max IoU
(b)

Fig. 1: Schematic of proposed event detection procedure. (a) Input data $\mathbf{x}$ is fed to the model f, which outputs predictions for event classes and localizations for each default event in ε^d. (b) The IoU for each predicted ε_j^* is then calculated with respect to the true event ε_i and non-maximum suppression is applied to match up true events and predictions. In the current case, the predicted event marked in black has the highest IoU with the true event in blue. For more information, see [16], [19].

B. Detection model overview

In brief, the proposed model receives as input a tensor $\mathbf{x} \in \mathbf{R}^{C \times T}$ containing C channels of data in a segment of T samples, along with a set of events $\{\varepsilon_i \in \mathbf{R}^2 \mid \varepsilon_i = (\varrho_i, \delta_i)\,, i = 1, \ldots, N_{\mathbf{x}}\}$, were $N_{\mathbf{x}}$ is the number of events in the associated time segment and (ϱ_i, δ_i) are the start time and duration of event ε_i. The objective of the deep learning model f is then to infer $\{\varepsilon_i\}$ given $\mathbf{x}$. To do this, a set of default events $\{\varepsilon_j^d \in \mathbf{R}^2 \mid j = 1, \ldots, N_d,\, N_d = T/\tau\}$ is generated over the segment of T samples, where τ is the size of each default event window in samples. The model outputs

TABLE II: Proposed network architecture. ϕ_C, linear mixing module; ϕ_T, temporal feature extraction module; ϕ_R, recurrent neural network module; ψ_{clf}, event classification module; ψ_{loc}, event localization module; *C*, number of input channels; *T*, number of samples in segments; $\tilde{C} = 2^{2+n_{\max}}$, number of output channels; *K*, number of event classes; N_d, number of default events in segment; $\tilde{T} = {}^{T}/_{2^{n_{\max}}}$, reduced temporal dimension; bGRU, bidirection gated recurrent unit; ReLU, rectified linear unit.

Module	Input dim.	Output dim.	Type	Kernel size	No. kernels	Stride	Activation
ϕ_C	(C, T)	(C, T)	1D convolution	C	C	1	linear
	(C, T)	$(8, T)$	1D convolution	3	8	1	–
$\phi_{T,\text{init}}$	$(8, T)$	$(8, T)$	Batch norm.	–	8	–	ReLU
	$(8, T)$	$(8, {}^{T}/_{2})$	1D max. pool.	2	–	2	–
	$\left(2^{n+1}, {}^{T}/_{2^{n\ \ 1}}\right)$	$\left(2^{n+2}, {}^{T}/_{2^{n\ \ 1}}\right)$	1D convolution	3	2^{n+2}	1	–
$\phi_{T,k}$ $n = 2, \ldots, n_{\max}$	$\left(2^{n+2}, {}^{T}/_{2^{n\ \ 1}}\right)$	$\left(2^{n+2}, {}^{T}/_{2^{n\ \ 1}}\right)$	Batch norm.	–	2^{n+2}	–	ReLU
	$\left(2^{n+2}, {}^{T}/_{2^{n\ \ 1}}\right)$	$\left(2^{n+2}, {}^{T}/_{2^{n}}\right)$	1D max. pool.	2	–	2	–
ϕ_R	$(\tilde{C}, \tilde{T})$	$(2 \times \tilde{C}, \tilde{T})$	bGRU	$\tilde{C}$	–	–	–
ψ_{clf}	$(\tilde{C}, \tilde{T})$	$((K+1)\, N_d, 1)$	1D convolution	$\tilde{T}$	$(K+1)\, N_d$	$\tilde{T}$	softmax for each $K + 1$ kernel
ψ_{loc}	$(\tilde{C}, \tilde{T})$	$(2N_d, 1)$	1D convolution	$\tilde{T}$	$2N$	$\tilde{T}$	linear

probabilities for K classes including the default, non-event class for each default event window. The probability for a given class k in the default event window ε_j^d must be greater than a classification threshold θ_{clf}. In order to select among many possible candidates of predicted events, all predicted events of class k over the possible events in N_d is subjected to non-maximum suppression using the intersection-over-union (IoU, Jaccard index) as in [18]. A high-level schematic of the detection model is shown in Fig. 1.

C. Network architecture

The architecture for the proposed PSG event detection model follows closely the event detection algorithms described in [16], [17], albeit with some specific changes. An overview of the proposed network in the model *f* is provided in Table II. Briefly, the model comprises three modules:

1) a channel mixing module $\phi_C : \mathbf{R}^{C\times T} \to \mathbf{R}^{C\times T}$;
2) a feature extraction module $\phi_T : \mathbf{R}^{C\times T} \to \mathbf{R}^{\tilde{C}\times\tilde{T}}$;
3) and an event detection module ψ,

the latter containing two submodules performing event classification $\psi_{\text{clf}} : \mathbf{R}^{\tilde{C}\times\tilde{T}} \to \mathbf{R}^{(K+1)\times N_d}$ and event localization $\psi_{\text{loc}} : \mathbf{R}^{\tilde{C}\times\tilde{T}} \to \mathbf{R}^{2\times N_d}$, respectively. The difference between these two submodules is that ϕ_{clf} outputs the probability of the default, non-event class and K event classes, while ϕ_{loc} predicts a start time and a duration of all predicted events relative to a specific default event window. The channel mixing module ϕ_C receives a segment of input data $x \in \mathbf{R}^{C\times T}$, where C is the number of input channels and T is the number of time samples in the given segment, and subsequently performs linear channel mixing using 1D convolutions to synthesize C new channels. Following ϕ_C, the feature extraction module ϕ_T consists of $n_{\max}$ blocks with the first block $\phi_{T,1} : \mathbf{R}^{C\times T} \to \mathbf{R}^{8\times {}^{T}/_{2}}$ and the nth block $\phi_{T,n} : \mathbf{R}^{2^{n+1}\times {}^{T}/_{2^{n\ \ 1}}} \to \mathbf{R}^{2^{k+2}\times {}^{T}/_{2^{n}}}$. All $n_{\max}$ blocks implement $\phi_{T,n}$ using 1D convolution layers followed by batch normalization of the feature maps, rectified linear unit activation, and final 1D maximum pooling layers across the temporal dimension. Kernel sizes and strides for convolution and max. pool. layers in ϕ_T were set to 3 and 1, and 2 and 2, respectively, while the number of feature maps in $\phi_{T,n}$ was set to 2^{n+2}. The event classification submodule ψ_{clf} is implemented a 1D convolution layer across the entire data volume using $(K+1)N_d$ feature maps of size and stride $\tilde{T} = T/2^{n_{\max}}$, where $K \in \mathbf{N}$ is the number of event classes to be detected and $N_d \in \mathbf{N}$ is the number of default event windows. The event localization submodule ψ_{loc} is likewise implemented using a 1D convolution layer across the entire data volume.

D. Data and event sampling

The proposed network requires an input tensor $x \in \mathbf{R}^{C\times T}$ containing PSG data in the time segment of size T as well as information about the associated events in the segment. Since the total number of segments in a standard PSG without any event data far outnumbers the number of segments with event data, we implemented a random sampling of non-event and event classes with the sampling probability of class k inversely proportional to the number of classes, such that $p_k = \frac{1}{K+1}$, $k = [0 \mathinner{..} K]$, where $k = 0$ is the default (non-event) class. At training step t, we thus sample a class k and afterwards randomly sample a single class k event ε_k between all class k events. Finally, we extract a segment of PSG data of size $C \times T$ with start of segment in the interval $[\bar{\varepsilon}_k - T, \bar{\varepsilon}_k + T]$, where $\bar{\varepsilon}_k$ is the sample midpoint of ε_k. This ensures that each $\mathbf{x}$ overlaps 50% with at least one associated event.

E. Optimization of network parameters

The network parameters were optimized using mini-batch stochastic gradient descent with initial learning rate of 10^{-3} and a momentum of 0.9. Minibatches were balanced with respect to the detected classes. The optimization was performed with respect to the same loss function described in [16], [17] and the network was trained until convergence determined by no decrease in the loss on the EVAL set over 10 epochs of TRAIN data. We also employed learning rate decay with a factor of 2 every 5 epochs of non-decreasing EVAL loss.

F. Experimental setups

In this study, we examined two different experimental setups.

a) Experiment A: First, we investigated the differences in predictive performance using a static vs. a dynamic default event window size. This was realized by running six separate training runs with $\tau \in \{3, 5, 10, 15, 20, 30\} \times f_s$, as well as a single training run where f was evaluated for all $\{3, 5, 10, 15, 20, 30\} \times f_s$. The best performing model was determined by evaluating F1 score on the EVAL set for both LM and AR detection.

b) Experiment B: Second, we tested a network where we added a recurrent processing block ϕ_R after the feature extraction block ϕ_T as shown in grey in Table II. We considered a single bidirectional gated recurrent unit (bGRU) layer with $\tilde{C}$ units. Predictions were evaulated across multiple time-scales $\tau \in \{3, 5, 10, 15\} \times f_s$

All experiments were implemented in PyTorch 1.0 [25].

G. Performance metrics

All models were evaluated on the EVAL and TEST sets using precision (Pr), recall (Re), and F1 scores (F1):

$$\mathrm{Pr} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{Re} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{F1} = 2\frac{\mathrm{Pr} * \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}},$$

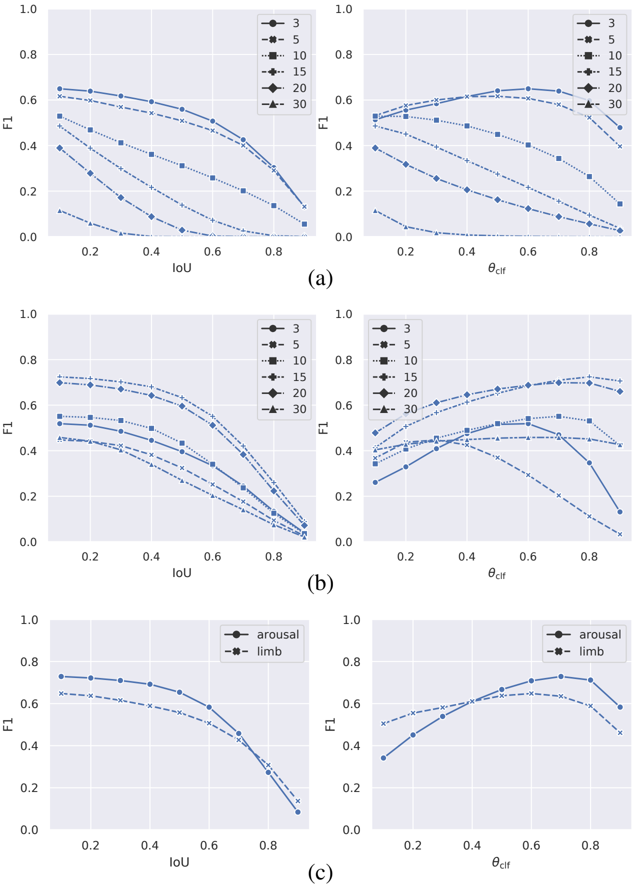
where TP, FP, and FN, are the number of true positives, false positives and false negatives, respectively.

H. Statistical analysis

Demographic and polysomnographic variables were tested for subset differences with Kruskall-Wallis H-test for independent samples.

IV. RESULTS AND DISCUSSION

Shown in Figs. 2a and 2b are the F1 scores as a function of IoU and the classification threshold θ_{clf} for both the LM and AR detection models. It is apparent that both models perform best with a minimum overlap (IoU $= 0.1$) with their respective annotated events, and do not benefit from increasing the overlap. This might be caused by the fact that the annotated events might not be precise enough, and not due to issues with the model itself. For example, it is not uncommon to only mark the beginning of an event in standard sleep scoring software, as the duration will automatically be annotated by a default length, such as 3 s for ARs, and 0.5 s for LM (which is the minimum duration as defined by the AASM guidelines [1]). Future studies will be able to confirm this by either collecting a precisely annotated cohort, or by investigating the average start time and duration discrepancies between annotated and predicted events.



(a)

(b)

(c)

Fig. 2: Experiment A: Optimizing IoU and θ_{clf} in static models on the EVAL set by varying default event window size in seconds in $\{3, 5, 10, 15, 20, 30\}$ (a)-(b). Left panels show the IoU vs. F1 score, while right panels show classification threshold θ_{clf} against F1 score. (a) LM model. Here, the model performs best for IoU $= 0.1$ and $\theta_{\mathrm{clf}} = 0.6$ using a window size of $\tau = 3\,\mathrm{s} \times f_s$. (b) AR model. Here, the model performs best for IoU $= 0.1$ and $\theta_{\mathrm{clf}} = 0.8$ using a window size of $\tau = 15\,\mathrm{s} \times f_s$. (c) Dynamic models show optimal performance for IoU $= 0.1$ and $\theta_{\mathrm{clf}} = 0.7$ and $\theta_{\mathrm{clf}} = 0.6$ for AR and LM detection, respectively.

It is also apparent from Figs. 2a and 2b that both detection models benefit from imposing a strict classification threshold. Specifically, LM detection performance as measured by F1 was highest with $\theta_{\mathrm{clf}} = 0.6$, while maximum AR detection performance was attained with an even higher θ_{clf} of 0.8.

Furthermore, we explored allowing for multiple time-scales in the dynamic models, shown in Fig. 2c. It was hypothesized that having the default event windows dynamic instead of static would allow for more flexibility and thus better predictive performance, however, we observed no significant differences between the optimal static window and the dynamic window model.

Shown in Fig. 3 are the performance curves for the RNN (bidirectional GRU) version of the proposed model for each of the two event detection tasks. While the optimal IoU and

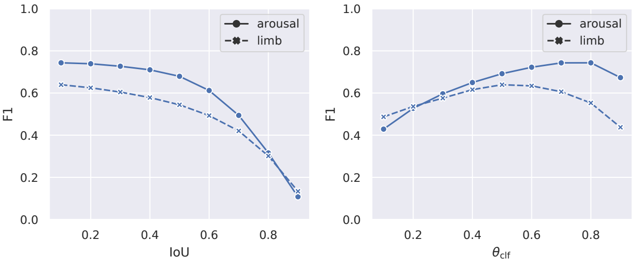


Fig. 3: Experiment B. F1 performance on the EVAL set as a function of IoU and θ_{clf} for AR and LM detection when adding the ϕ_R module. Best performance is seen for IoU $=$ 0.1 for both AR and LM detection, and $\theta_{\mathrm{clf}} = 0.6$ and $\theta_{\mathrm{clf}} =$ 0.8 for LM and AR detection, respectively.

TABLE III: Application of optimized models on TEST data. Data are shown as subject-averaged F1, precision (Pr) and recall (Re) with associated standard deviations. Top four rows correspond to Experiment A, while bottom two rows correspond to Experiment B. AR: arousal; LM: leg movement; RNN: recurrent neural network.

Model	F1	Pr	Re
LM, static	0.648 ± 0.148	0.631 ± 0.181	0.720 ± 0.141
AR, static	0.727 ± 0.102	0.706 ± 0.113	0.771 ± 0.132
LM, dynamic	0.647 ± 0.148	0.627 ± 0.181	0.722 ± 0.14
AR, dynamic.	0.729 ± 0.102	0.699 ± 0.115	0.785 ± 0.131
LM, RNN	0.639 ± 0.147	0.606 ± 0.180	0.727 ± 0.126
AR, RNN	0.749 ± 0.105	0.772 ± 0.107	0.748 ± 0.138

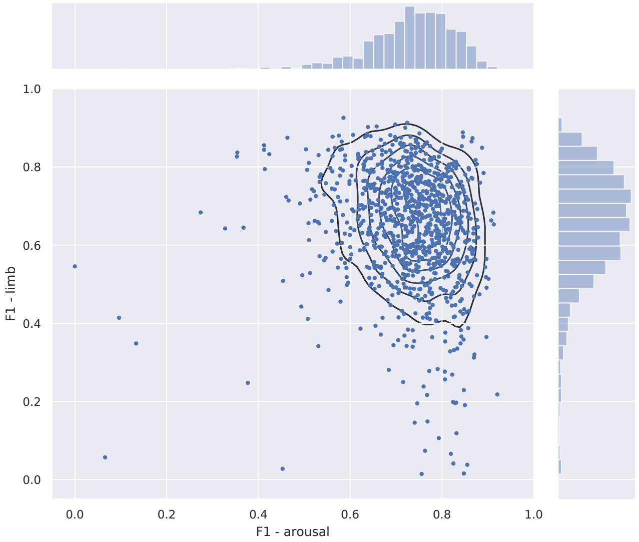


Fig. 4: Visualization of F1 scores for both AR and LM detection using the dynamic model.

θ_{clf} points are unchanged from the static/dynamic models presented in Fig. 2, the optimal F1 value for AR detection is increased by incorporating temporal dependencies in the model. The reverse is true for LM detection, which saw a slight decrease in predictive performance caused by a lower precision (see Table III). Future work should consider optimizing predictive performance by investigating the effects of varying the number of bGRU layers and the number of hidden units in ϕ_R, since this was not performed here.

Application of the optimal models on the TEST data is shown in Table III. We observed that with the given architecture of f and the given labels and input data in TRAIN, LM detection was maximal for the model with a static/dynamic window, while adding a recurrent module only positively impacted AR prediction. We observed a general decrease in both precision and recall for LM detection when adding ϕ_R, while precision actually increased and recall decreased for AR detection. An example visualization of the joint distribution of F1 scores obtained from the dynamic model applied to the TEST data is shown in Fig. 4. While some outliers are readily observable especially for LM detection, the majority of subject F1 scores follows an approximate bivariate normal distribution.

Subset partitions were reasonably well-distributed with no significant differences between key variables, see Table I. An exception is the AHI, although the associated effect is small and most likely a result of the low sample size in EVAL compared to TRAIN and TEST. It is noted, that although AHI, AI, and PLMI are not normally distributed and summarizing these variables with standard deviations is invalid, it is nevertheless standard practice in sleep medicine and thus presented the same way here. We performed little data cleaning in order to provide as much data and variation to the deep learning model as possible, however, future efforts should explore and apply inclusion criteria such as minimal total sleep time, artifact detection and removal of studies with severe artifacts. We did impose a trivial lower bound on the number of scored events (>0) for a PSG to be included in this study, but stricter requirements could potentially improve model performance.

In this work, we investigated 'systemic' PSG events present in multiple signal modalities instead of EEG-specific events, which required changes to the network architecture. Specifically, we kept the signal modality encoded in the first dimension of the tensor propagated through the network, which allowed for the use of one-dimensional convolutional operators. By performing 1D convolutions and keeping the channel information in the feature maps instead of keeping them as separate dimensions and performing 2D convolutions as proposed in [16], [17], we simplify and reduce the number of computations and training time by a factor $\propto C$. However, we did not investigate the effects of modeling the conditional probability of AR and LM occurrence, but the proposed architecture is versatile enough to detect both events jointly as well as separately. Previous work also suggest that detecting multiple objects at the same time is of high interest and leads to (at least) non-inferior performances [16]–[19].

Additionally, we speculated that the temporal dynamics of the PSG signals were important for optimal event detection performance. Although the effects were small, we did show an increase in F1 score in AR detection when adding an RNN module to the network before the detection module.

However, this was not the case for LM detection, which is most likely due to the different temporal and physiological characteristics of the two events in question.

Future efforts will be addressing the fact that in the current modeling scheme, events are mutually exclusive given a certain default event window size. However, it is common to see ARs and LMs as a result of one another, and thus, if the window size is too small, a more unlikely event as measured by classification threshold and IoU will be removed even if it matches up to a specific true event of a certain class.

V. CONCLUSIONS

We have proposed a deep learning model that extends on previous work and shows promise in automatic detection of arousals and leg movements during sleep. The proposed model is flexible in allowing for the detection of multiple events of distinct physiological natures. Future work will expand further on adding more signals and event classes in order to complete a general purpose sleep analysis tool.

ACKNOWLEDGMENT

Some of the computing for this project was performed on the Sherlock cluster. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results.

REFERENCES

[1] 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. Darien, Il: American Academy of Sleep Medicine, 2018.

[2] P. Halász, M. Terzano, L. Parrino, and R. Bódizs, "The nature of arousal in sleep," *J. Sleep Res.*, vol. 13, no. 1, pp. 1–23, 2004.

[3] L. J. Findley, M. E. Unverzagt, and P. M. Suratt, "Automobile Accidents Involving Patients with Obstructive Sleep Apnea," *Am. Rev. Respir. Dis.*, vol. 138, no. 2, pp. 337–340, 1988.

[4] R. Ferri, B. B. Koo, D. L. Picchietti, and S. Fulda, "Periodic leg movements during sleep: phenotype, neurophysiology, and clinical significance," *Sleep Med.*, vol. 31, pp. 29–38, 2017.

[5] American Academy of Sleep Medicine, *International classification of sleep disorders*, 3rd ed. Darien, Il: American Academy of Sleep Medicine, 2014.

[6] 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. 901–8, 2000.

[7] R. S. Rosenberg and S. Van Hout, "The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring," *J. Clin. Sleep Med.*, vol. 9, no. 1, pp. 81–87, 2013.

[8] M. Younes, J. Raneri, and P. Hanly, "Staging sleep in polysomnograms: Analysis of inter-scorer variability," *J. Clin. Sleep Med.*, vol. 12, no. 6, pp. 885–894, 2016.

[9] M. H. Bonnet, K. Doghramji, T. Roehrs, E. J. Stepanski, S. H. Sheldon, A. S. Walters, M. Wise, and A. L. Chesson Jr, "The scoring of arousal in sleep: reliability, validity, and alternatives," *J. Clin. Sleep Med.*, vol. 3, no. 2, pp. 133–145, 2007.

[10] 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.

[11] A. N. Olesen, P. Jennum, P. Peppard, E. Mignot, and H. B. D. Sorensen, "Deep residual networks for automatic sleep stage classification of raw polysomnographic waveforms," in *2018 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Honolulu, HI, USA, July 18 – 20, 2018, pp. 3713–3716.

[12] J. B. Stephansen, A. N. Olesen, M. Olsen, A. Ambati, E. B. Leary, H. E. Moore, O. Carrillo, L. Lin, F. Han, H. Yan, Y. L. Sun, Y. Dauvilliers, S. Scholz, L. Barateau, B. Hogl, A. Stefani, S. C. Hong, T. W. Kim, F. Pizza, G. Plazzi, S. Vandi, E. Antelmi, D. Perrin, S. T. Kuna, P. K. Schweitzer, C. Kushida, P. E. Peppard, H. B. D. Sorensen, P. Jennum, and E. Mignot, "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy," *Nat. Commun.*, vol. 9, no. 1, p. 5229, 2018.

[13] 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, 2018.

[14] S. Biswal, H. Sun, B. Goparaju, M. B. Westover, J. Sun, and M. T. Bianchi, "Expert-level sleep scoring with deep neural networks," *J. Am. Med. Informatics Assoc.*, vol. 25, no. 12, pp. 1643–1650, 2018.

[15] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, 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.*, 2019.

[16] S. Chambon, V. Thorey, P. J. Arnal, E. Mignot, and A. Gramfort, "A deep learning architecture to detect events in EEG signals during sleep," in *2018 IEEE Int. Work. Mach. Learn. Signal Process. (MLSP)*, Aalborg, Denmark, Sept. 17 – 20, 2018, pp. 1–6.

[17] S. Chambon, V. Thorey, P. Arnal, E. Mignot, and A. Gramfort, "DOSED: a deep learning approach to detect multiple sleep micro-events in EEG signal," *J. Neurosci. Methods*, 2019, https://doi.org/10.1016/j.jneumeth.2019.03.017.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 779–788.

[19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," in *Computer Vision (ECCV)*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 21–37.

[20] J. B. Blank, P. M. Cawthon, M. L. Carrion-Petersen, L. Harper, J. P. Johnson, E. Mitson, and R. R. Delay, "Overview of recruitment for the osteoporotic fractures in men study (MrOS)," *Contemp. Clin. Trials*, vol. 26, no. 5, pp. 557–568, 2005.

[21] E. Orwoll, J. B. Blank, E. Barrett-Connor, J. Cauley, S. Cummings, K. Ensrud, C. Lewis, P. M. Cawthon, R. Marcus, L. M. Marshall, J. McGowan, K. Phipps, S. Sherman, M. L. Stefanick, and K. Stone, "Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study – A large observational study of the determinants of fracture in older men," *Contemp. Clin. Trials*, vol. 26, no. 5, pp. 569–585, 2005.

[22] T. Blackwell, K. Yaffe, S. Ancoli-Israel, S. Redline, K. E. Ensrud, M. L. Stefanick, A. Laffan, and K. L. Stone, "Associations Between Sleep Architecture and Sleep-Disordered Breathing and Cognition in Older Community-Dwelling Men: The Osteoporotic Fractures in Men Sleep Study," *J. Am. Geriatr. Soc.*, vol. 59, no. 12, pp. 2217–2225, 2011.

[23] D. A. Dean, A. L. Goldberger, R. Mueller, M. Kim, M. Rueschman, D. Mobley, S. S. Sahoo, C. P. Jayapandian, L. Cui, M. G. Morrical, S. Surovec, G.-Q. Zhang, and S. Redline, "Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource," *Sleep*, vol. 39, no. 5, pp. 1151–1164, 2016.

[24] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline, "The National Sleep Research Resource: towards a sleep data commons," *J. Am. Med. Informatics Assoc.*, vol. 25, no. 10, pp. 1351–1358, 2018.

[25] A. Paszke, G. Chanan, Z. Lin, S. Gross, E. Yang, L. Antiga, and Z. Devito, "Automatic differentiation in PyTorch," Long Beach, CA, USA, 2017.

E

PAPER V

TITLE: Deep transfer learning for improving single-EEG arousal detection

AUTHORS: Alexander Neergaard Olesen, Poul Jennum, Emmanuel Mignot, and Helge B. D. Sorensen

CONFERENCE: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)

STATUS: Accepted for publication

DOI:

FULL CITATION: A. N. Olesen, P. Jennum, E. Mignot and H. B. D. Sorensen, "Deep transfer learning for improving single-EEG arousal detection,", 2020, *accepted*

COPYRIGHT INFORMATION:

Deep transfer learning for improving single-EEG arousal detection

Alexander Neergaard Olesen*,[1,2,3], *Member, IEEE*, Poul Jennum[3],
Emmanuel Mignot[2] and Helge B. D. Sorensen[1], *Senior Member, IEEE*

***Abstract*— Datasets in sleep science present challenges for machine learning algorithms due to differences in recording setups across clinics. We investigate two deep transfer learning strategies for overcoming the channel mismatch problem for cases where two datasets do not contain exactly the same setup leading to degraded performance in single-EEG models. Specifically, we train a baseline model on multivariate polysomnography data and subsequently replace the first two layers to prepare the architecture for single-channel electroencephalography data. Using a fine-tuning strategy, our model yields similar performance to the baseline model (F1=0.682 and F1=0.694, respectively), and was significantly better than a comparable single-channel model. Our results are promising for researchers working with small databases who wish to use deep learning models pre-trained on larger databases.**

I. INTRODUCTION

A principal tool in the analysis of sleep is the polysomnography (PSG). Standard PSGs contain electroencephalography (EEG), electrooculography (EOG), electromyography (EMG) from below the chin and lower limbs, electrocardiography, respiratory effort, and blood oxygenation, which is manually analysed by sleep experts according to guidelines published by the American Academy of Sleep Medicine [1].

Experts score sleep stages and annotate discrete events, such as arousals (short awakenings during sleep, $\leq 15\,\mathrm{s}$), limb movements, and decreased respiratory effort characterized by apneas (complete cessation of breathing), hypopneas (partial cessation of breathing), and desaturations (decreases in oxygen desaturation). Low inter-rater reliability has been reported for sleep stages scoring in multiple studies [2]–[4], arousals [5], and respiratory events [6], [7], prompting extensive research in automated methods for sleep analysis [8]–[14].

Designing reliable and robust systems for automated sleep analysis based on machine learning algorithms often require multiple heterogenous data sources of sufficient size. However, due to differences in clinical practice, very few datasets in sleep science are standardized with regards to recording setups despite guidelines from the AASM.

In these cases, we end up with a *channel mismatch problem*, in which the overlap between our source and target domains is small, and the domains are possibly disjointed. Recent studies have investigated the use of deep transfer learning to solve the channel mismatch problem when training and testing sleep stage classification models [15], [16]. The authors found that using a fine-tuning strategy significantly improved the performance of sleep stage scoring models when trained on various combinations of EEG and EOG channels.

We present results on using deep transfer learning to address the channel mismatch problem, when the source and target domains differ both in the number and type of channel modalities. Specifically, our source domain consists of multivariate PSG data comprising left and right central EEG, left and right EOG, and submental EMG recordings, while our target domain consists of only a single central EEG channel. We show that by employing a simple fine-tuning strategy on a pre-trained network stripped of the initial two layers, we can effectively reach the same level of F1 score as when using the full set of PSG data.

II. METHODS

Notation: We denote by $\llbracket a, b \rrbracket$ the set of integers $\{n \in \mathbb{N} \mid a \leq n \leq b\}$ with $\llbracket N \rrbracket$ being shorthand for $\llbracket 1, N \rrbracket$, and by $n \in \llbracket N \rrbracket$ the nth sample in $\llbracket N \rrbracket$. A model for a given experiment is denoted by $\mathcal{M}_{(\cdot)}$, while an optimized model is superscripted with a star as $\mathcal{M}^*_{(\cdot)}$. A segment of PSG data is denoted by $\mathbf{x} \in \mathbb{R}^{C \times T}$, where C, T is the number of channels and the duration of the segment in samples, respectively.

A. Data

We collected PSGs from 1500 subjects in the MrOS Sleep Study [17]–[19] from the National Sleep Research Resource repository [20], [21]. From each PSG, we extracted left and right EEG, left and right EOG, and chin EMG. EEG and EOG channels were referenced to the contralateral mastoid process. For each PSG, we also extracted time-stamped arousal scorings containing starts and durations of scored arousal events. We did not exclude any PSGs from this study based on sleep duration, number of arousal events, or similar criteria.

B. Data partitioning

The 1500 PSGs were initially partitioned into three subsets TRAIN_1, EVAL_1, and TEST_1 containing 400, 100 and 1000 PSGs, respectively. Furthermore, we additionally partitioned TEST_1 into three smaller subsets TRAIN_2, EVAL_2, and TEST_2 containing 400, 100, and 500 PSGs, respectively.

Research supported by the Klarman Family Foundation, Technical University of Denmark, and University of Copenhagen with supporting grants from Reinholdt W. Jorck & Wife's Foundation, Knud Højgaard Foundation, Otto Mønsted Foundation, Vera & Carl Michaelsens Foundation, Augustinus Foundation, and Stibo Foundation.

†Corresponding author: aneol@dtu.dk

[1] Department of Health Technology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark.

[2] Center for Sleep Sciences and Medicine, Stanford University, Palo Alto, CA 94304, USA.

[3] Danish Center for Sleep Medicine, University Hospital Copenhagen, 2600 Glostrup, Denmark

TABLE I

Network architecture overview.

Module	Layer type	Kernel	Stride	Feature maps	Input size	Output size	Activation
$\mathbf{x}$	Input	—	—	—	$C \times T$	$1 \times C \times T$	—
ϕ_{mix}	2D convolution	$(C, 1)$	$(1, 1)$	C	$1 \times C \times T$	$C \times 1 \times T$	ReLU
$\varphi_{\text{conv},1}$	2D convolution	$(1, c)$	$(1, s)$	$2f_0$	$C \times 1 \times T$	$2f_0 \times 1 \times T/s$	—
	Batch norm.	—	—	$2f_0$	$2f_0 \times 1 \times T/s$	$2f_0 \times 1 \times T/s$	ReLU
$\varphi_{\text{conv},k}$	2D convolution	$(1, c)$	$(1, s)$	$f_0 2^k$	$f_0 2^{k-1} \times 1 \times T/s^{k-1}$	$f_0 2^k \times 1 \times T/s^k$	—
$k \in [\![2, k_{\max}]\!]$	Batch norm.	—	—	$f_0 2^k$	$f_0 2^k \times 1 \times T/s^k$	$f_0 2^k \times 1 \times T/s^k$	ReLU
φ_{rec}	bGRU	—	—	f'	$f' \times 1 \times T'$	$f' \times 2 \times T'$	—
ψ_{clf}	2D convolution	$(2, 1)$	$(1, 1)$	$(K+1)N_d$	$f' \times 2 \times T'$	$(K+1)N_d \times 1 \times T'$	Softmax over $K+1$
ψ_{loc}	2D convolution	$(2, 1)$	$(1, 1)$	$2N_d$	$f' \times 2 \times T'$	$2N_d \times 1 \times T'$	Linear
$\mathbf{z}$	Output, $\mathbf{p}$	—	—	—	$(K+1)N_d \times 1 \times T'$	$N_d \times T' \times (K+1)$	—
	Output, $\mathbf{y}$	—	—	—	$2N_d \times 1 \times T'$	$N_d \times T' \times 2$	—

$\mathbf{x}$, input containing PSG data; $\mathbf{z}$, output containing predicted arousal probabilities and associated start and duration predictions; ϕ_{mix}, non-linear mixing block; φ_{conv}, convolutional feature extraction block; φ_{rec} recurrent feature extraction block; ψ_{clf}, event classification block; ψ_{loc}, event localization block; C, number of input channels; T, number of samples in a segment of PSG data; c, temporal kernel size; s, temporal stride; f_0, base number of feature maps; $f' = f_0 2^{k_{\max}}$, maximum number of feature maps; $T' = T/s^{k_{\max}}$, reduced temporal dimension in samples; N_d, number of default event windows in segment; K, number of classes; ReLU, rectified linear unit; bGRU, bidirectional gated recurrent unit.

C. Preprocessing pipeline

All signals were resampled to 128 Hz using poly-phase filtering with a Kaiser window ($\beta = 5.0$) prior to subsequent processing. Extracted EEG and EOG signals were filtered with 2nd order Butterworth IIR bandpass filters with cutoff frequencies 0.3 Hz and 35 Hz. Chin EMG was filtered with a 4th order Butterworth IIR highpass filter with a cutoff frequency of 10 Hz. Filtered signals were subsequently standardized by

$$\mathbf{x}^{(i)} = \frac{\tilde{\mathbf{x}}^{(i)} - \boldsymbol{\mu}^{(i)}}{\boldsymbol{\sigma}^{(i)}}, \tag{1}$$

where $\tilde{\mathbf{x}}^{(i)} \in \mathbb{R}^{C \times T}$ is the raw matrix containing C input channels and T samples, and $\boldsymbol{\mu}^{(i)}, \boldsymbol{\sigma}^{(i)} \in \mathbb{R}^{C}$ are the mean and standard deviation vectors for the i'th PSG, respectively.

D. Model setup

We expand upon previous work using similar models for sleep event detection [22]–[24]. Briefly, the model takes as input a tensor of PSG data $\mathbf{x} \in \mathbb{R}^{C \times T}$ and outputs

$$\mathbf{z} = (\mathbf{p}, \mathbf{y}) \in \mathbb{R}^{N_d \times T' \times (K+1)} \times \mathbb{R}^{N_d \times T' \times 2} \tag{2}$$

containing predicted arousal probabilities $\mathbf{p}$ and associated start and durations for predicted arousal events $\mathbf{y}$. The differentiable function underlying the model comprises a deep neural network architecture consisting of the following modules:

a) Input mixing module: Here, non-linear combinations of the input PSG data $\mathbf{x}$ are made using a non-linear mixing block $\phi_{\text{mix}} : \mathbb{R}^{1 \times C \times T} \to \mathbb{R}^{C \times 1 \times T}$.

b) Feature extraction module: This module contains two components. The first is a convolutional feature extraction block $\varphi_{\text{conv}} : \mathbb{R}^{C \times 1 \times T} \to \mathbb{R}^{f' \times 1 \times T'}$ consisting of $k_{\max}$ successions of convolutional, batch normalization, and rectified linear unit (ReLU) layers. Second is a recurrent feature extraction block $\varphi_{\text{rec}} : \mathbb{R}^{f' \times 1 \times T'} \to \mathbb{R}^{f' \times 2 \times T'}$ with $f' = f_0 2^{k_{\max}}$ hidden units. The φ_{conv} block is responsible for bulk feature extraction and temporal decimation using strided convolutions, while φ_{rec} processes the raw features across the reduced temporal dimension using a bidirectional gated recurrent unit [25] with f' hidden units.

c) Event detection module: The output from φ_{rec} is processed by two separate blocks: $\psi_{\text{clf}} : \mathbb{R}^{f' \times 2 \times T'} \to \mathbb{R}^{(K+1)N_d \times 1 \times T'}$ outputs the tensor $\mathbf{p}$ containing predicted arousal probabilities for each time point $t \in [\![T']\!]$ for each default event window. $\psi_{\text{loc}} : \mathbb{R}^{f' \times 2 \times T'} \to \mathbb{R}^{2N_d \times T'}$ outputs the tensor $\mathbf{y}$ containing predicted start time and durations of arousal events. Both ψ_{clf} and ψ_{loc} are implemented using $(2, 1)$ convolutions rather than convolutions over the entire volume as in [22]–[24]. This serves a dual purpose: the first is to reduce the number of parameters to make the network more memory-efficient, while the second purpose is to allow the kernel and feature maps to be temporally invariant.

For a detailed description of the network architecture, see Table I.

E. Loss objective

The network parameters were optimized according to a three-component loss objective comprising a localization loss ℓ_{loc} and a positive and negative classification loss ℓ_+ and ℓ_-, respectively, such that

$$\ell = \ell_{\text{loc}} + \ell_+ + \ell_-. \tag{3}$$

The localization loss was calculated using a Huber function

$$\ell_{\text{loc}} = \frac{1}{N_{\pi \setminus \emptyset}} \sum_{i \in \pi \setminus \emptyset} h^{(i)} \tag{4}$$

$$\mathbf{h} = \begin{cases} 0.5(\mathbf{y}-\mathbf{t})^2, & \text{if } |\mathbf{y}-\mathbf{t}| < 1, \\ |\mathbf{y}-\mathbf{t}| - 0.5, & \text{otherwise,} \end{cases} \tag{5}$$

where $i \in \pi\backslash\emptyset$ indicates event windows with a non-empty arousal target. Contributions from the positive/negative classification losses were calculated using a focal loss function [26]:

$$\ell_+ = \frac{1}{N_{\pi\backslash\emptyset}} \sum_{i\in\pi\backslash\emptyset} -\alpha\left(1-\mathbf{p}\right)^{\gamma} \log\left(\mathbf{p}\right), \text{ and} \tag{6}$$

$$\ell_- = \frac{1}{N_{\pi=\emptyset}} \sum_{i\in\pi=\emptyset} -\alpha\left(1-\mathbf{p}\right)^{\gamma} \log\left(\mathbf{p}\right), \tag{7}$$

where $\alpha = 0.25$ and $\gamma = 2$. This serves to counter the class imbalance in a single data segment, which typically consists of many event windows with few positive examples.

F. Experimental setups

We investigated the channel mismatch problem with the following four experimental setups:

a) Full montage baseline (FM): In this experiment, we trained the event detection algorithm on TRAIN_1 using $C = 5$ channels: left/right central EEG, left/right EOG, and chin EMG. Convergence and the optimal detection threshold were assessed on EVAL_1 and performance was evaluated on TEST_2. The optimal baseline model was used as an initialization for the two transfer learning experiments described below.

b) Pretraining (PT): The optimal model $\mathcal{M}^*_{\text{FM}}$ was used in this experiment as an initialization for $\mathcal{M}_{\text{PT}}$. We adjusted the mixing module and first convolutional layer in the feature extraction module to account for the channel mismatch by replacing the convolutional and batch normalization layers, and subsequently trained these from scratch. The rest of the weights and bias terms were frozen to the optimized values from $\mathcal{M}^*_{\text{FM}}$. The network was trained on TRAIN_2 with only $C = 1$ channels (left central EEG, C3). Convergence and optimal detection threshold were assesed on EVAL_2, while final performance was evaluated on TEST_2.

c) Fine-tuning (FT): Similar to PT, the optimal model $\mathcal{M}^*_{\text{FM}}$ was used in this experiment as an initialization for $\mathcal{M}_{\text{FT}}$. Also, the mixing module and first convolutional layer in the feature extraction module were likewise adjusted. However, all other layers in $\mathcal{M}_{\text{FT}}$ were permitted to be further optmized by fine-tuning weights and bias terms during training. The model was trained using the same 400 PSGs from TRAIN_2 with the same $C = 1$ channel configuration as in PT.

d) Single EEG benchmark (SE): We benchmarked our two transfer learning experiments to a comparable situation in which an event detection model was trained on the same PSGs in TRAIN_2 using only the left central EEG (C3).

In all experimental runs, we optimized the loss objective in eq. (3) using the Adam optimization algorithm with a learning rate of $\alpha = 10^{-3}$ and the default parameter values $(\beta_1, \beta_2) = (0.9, 0.999)$ as suggested in [27]. We applied the same data sampling strategy as proposed in [24], in which a segment of data is sampled such that it contains at least 50% of a randomly sampled event across all PSGs. We used a default event window size of 15 s with 50 % overlap as this was found previously to work well for arousal detection [24].

All experiments were implemented in PyTorch 1.2 [28].

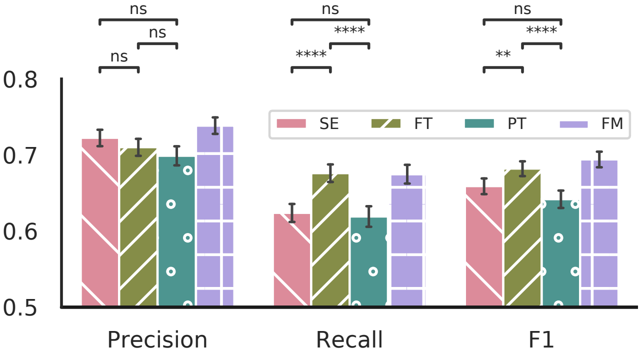


Fig. 1. Performance metrics as evaluated on TEST_2 for each experimental setup. Metrics are shown as means with 95% confidence interval as error bars. Note the y-axis scaling. SE: single-EEG. FT: fine-tuning. PT: pretraining. FM: full montage. ns: not significant, **: $p_{\text{adj}} \leq 10^{-2}$; ****: $p_{\text{adj}} \leq 10^{-4}$.

TABLE II
PERFORMANCE METRICS ACROSS EXPERIMENTS.

Experiment	**Precision**	**Recall**	**F1**
FM	0.739 ± 0.122	0.675 ± 0.139	0.694 ± 0.115
SE	0.723 ± 0.124	0.624 ± 0.137	0.659 ± 0.117
FT	$\mathbf{0.710 \pm 0.128}$	$\mathbf{0.676 \pm 0.130}$	$\mathbf{0.682 \pm 0.110}$
PT	0.699 ± 0.141	0.619 ± 0.153	0.642 ± 0.129

Metrics are shown evaluated on TEST_2 as means ± standard deviation. Best performing transfer learning experiment is shown in bold. SE: single-EEG. FT: fine-tuning. PT: pre-training. FM: full montage.

G. Performance evaluation

Bipartite matching were used to match detected and true events during training and testing. At test time, detected events were subjected to non-maximum suppression based on an intersection-over-union (IOU) of at least 0.5 between detected and true events. We evaluated the performance of our experimental setups using precision, recall and F1 scores.

H. Statistical analysis

We used Kruskal–Wallis one-way analysis of variance tests for differences in performance metrics between groups (SE, FT and PT) with a significance level of $\alpha = 0.05$. Post-hoc testing was performed with Mann-Whitney U-tests for each pair-combination (SE/FT, SE/PT, and FT/PT) likewise with $\alpha = 0.05$. We accounted for multiple comparisons by adjusting *p*-values with Bonferroni corrections.

III. RESULTS AND DISCUSSION

We show the results of the transfer learning experiments (FT, PT) as well as the baseline and benchmark experiments (FM, SE) in Fig. 1 and Table II. Performance metrics were not calculated for 10 subjects in TEST_2, as these did not have any scored arousals and are thus not reflected in Fig. 1 and Table II.

The baseline F1 performance is shown to be slightly lower than previously reported (0.694 ± 0.115 vs. 0.749 ± 0.105 [24]). However, our baseline model was trained on 400 subjects compared to 1485 in [24], which would account for the lower F1 score. By reducing the available input channels from $C = 5$ different modalities to $C = 1$ EEG channels as in the SE benchmark experiment, the F1 score drops to 0.659 ± 0.117, while the precision and recall scores likewise drop from 0.739 ± 0.122 to 0.723 ± 0.124, and 0.675 ± 0.139 to 0.624 ± 0.137, respectively.

We found statistically significant differences in F1 scores between SE, FT, and PT ($p = 3.189 \times 10^{-7}$). Post-hoc testing further revealed statistically significant differences between SE and FT ($p_{\text{adj}} = 2.224 \times 10^{-3}$), and FT and PT ($p_{\text{adj}} = 2.685 \times 10^{-7}$), but not between SE and PT ($p_{\text{adj}} = 0.080$). We also found that recall scores differed between experimental setups ($p = 7.085 \times 10^{-13}$). Post-hoc testing showed statistically significant differences between SE, FT ($p_{\text{adj}} = 5.180 \times 10^{-11}$), and FT and PT ($p_{\text{adj}} = 1.440 \times 10^{-9}$), but not between SE and PT ($p_{\text{adj}} = 1.000$). Lastly, we saw statistically significant differences in precision scores between experimental setups ($p = 0.033$), subsequent post-hoc testing did not reveal any statistical significant differences, when adjusting for multiple comparisons using the Bonferroni procedure (SE/FT, $p_{\text{adj}} = 0.214$; FT/PT, $p_{\text{adj}} = 1.000$; SE/PT, $p_{\text{adj}} = 0.037$).

Our results show, that for some scenarios, we can learn information present in multi-variate PSG data and efficitively transfer that information to a target domain containing only a single EEG channel. Specifically, the performance of our fine-tuning strategy is high enough that the mean F1 scores across subjects are statistically insifignicant, when comparing FT and FM setups (not shown).

Previous related work focused on the channel mismatch problem, when comparing different, but the same number of, channel modalities such as transferring EEG-based models to EOG-based target domains, and thus did not investigate how changing the model architecture might impact performance [15], [16]. In this work, we investigated transfer learning when the source and target domains only overlap by one input channel. This necessitates changing some parts of the underlying model architecture to accommodate the different number of input channels, and these changes might impact downstream feature extraction. We did not explore simply zeroing out a large number of input channels in this work, as this requires exhaustive search of which channel indices to zero out in the model based on the number of target input channels. Our strategy does not require this exhaustive search.

Our study applied a simple optimization strategy for the transfer learning experiments, which might limit the potential performance gain. This is especially relevant for the FT experiment. For example, one could experiment with with different learning rates and scheduling schemes for the initial layers and pre-trained layers, such that the initial layers were trained with a higher relative learning rate to compensate for their lack of initial training.

Furthermore, we explored transfer learning for the channel mismatch problem in a single cohort of patient recordings. Future directions of this research will investigate scenarios, where both the source and target domains, and the datasets are different.

IV. CONCLUSIONS

We show in our experiments that a simple fine-tuning strategy can be employed to transfer learning from a model based on multi-variate PSG data to a configuration where only a single EEG lead is available for detecting arousals, and that the difference between single-EEG and multivariate PSG performance is negligible. Future work will explore the effects of various combinations of datasets on the impact of generalized event detection, when the source and target domains do not overlap completely.

ACKNOWLEDGMENT

Some of the computing for this project was performed on the Sherlock cluster. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results.

We would also like to thank the National Sleep Research Resource team for their efforts in collecting, organizing and making available the PSG data used in this study.

REFERENCES

[1] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. L. Marcus, and B. V. Vaughn, *The AASM Manual for the Scoring of Sleep and Associated Events*, 2.6. Darien, IL: American Academy of Sleep Medicine, 2020.

[2] 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. 901–8, 2000.

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

[4] M. Younes, J. Raneri, and P. Hanly, "Staging sleep in polysomnograms: Analysis of inter-scorer variability," *J. Clin. Sleep Med.*, vol. 12, no. 6, pp. 885–894, 2016. DOI: 10.5664/jcsm.5894.

[5] M. H. Bonnet, K. Doghramji, T. Roehrs, *et al.*, "The Scoring of Arousal in Sleep: Reliability, Validity, and Alternatives," *J. Clin. Sleep Med.*, vol. 03, no. 02, pp. 133–145, 2007. DOI: 10.5664/jcsm.26815.

[6] U. J. Magalang, N.-H. Chen, P. A. Cistulli, *et al.*, "Agreement in the Scoring of Respiratory Events and Sleep Among International Sleep Centers," *Sleep*, vol. 36, no. 4, pp. 591–596, 2013. DOI: 10.5665/sleep.2552.

[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. 04, pp. 447–454, 2014. DOI: 10.5664/jcsm.3630.

[8] A. N. Olesen, P. Jennum, P. Peppard, E. Mignot, and H. B. D. Sorensen, "Deep residual networks for automatic sleep stage classification of raw polysomnographic waveforms," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Honolulu, HI, USA: IEEE, 2018, pp. 1–4. DOI: 10.1109/EMBC.2018.8513080.

[9] J. B. Stephansen, A. N. Olesen, M. Olsen, *et al.*, "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy," *Nature Communications*, vol. 9, no. 1, p. 5229, 2018. DOI: 10.1038/s41467-018-07229-3.

[10] 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, 2018. DOI: 10.1109/TNSRE.2018.2813138.

[11] S. Biswal, H. Sun, B. Goparaju, M. B. Westover, J. Sun, and M. T. Bianchi, "Expert-level sleep scoring with deep neural networks," *J. Am. Med. Informatics Assoc.*, vol. 25, no. 12, pp. 1643–1650, 2018. DOI: 10.1093/jamia/ocy131.

[12] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, 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.*, 2019. DOI: 10.1109/TNSRE.2019.2896659.

[13] L. Carvelli, A. N. Olesen, A. Brink-Kjær, *et al.*, "Design of a deep learning model for automatic scoring of periodic and non-periodicleg movements during sleep validated against multiple human experts," *Sleep Medicine*, DOI: 10.1016/j.sleep.2019.12.032.

[14] A. Brink-Kjaer, A. N. Olesen, P. E. Peppard, *et al.*, "Automatic Detection of Cortical Arousals in Sleep and their Contribution to Daytime Sleepiness," 2019. arXiv: 1906.01700 [q-bio.NC].

[15] H. Phan, O. Y. Chén, P. Koch, *et al.*, "Towards More Accurate Automatic Sleep Staging via Deep Transfer Learning," pp. 1–11, 2019. arXiv: 1907.13177 [cs.LG].

[16] H. Phan, O. Y. Chen, P. Koch, A. Mertins, and M. D. Vos, "Deep Transfer Learning for Single-Channel Automatic Sleep Staging with Channel Mismatch," in *2019 27th European Signal Processing Conference (EUSIPCO)*, A Coruña, Spain: IEEE, 2019, pp. 1–5. DOI: 10.23919/EUSIPCO.2019.8902977.

[17] J. B. Blank, P. M. Cawthon, M. L. Carrion-Petersen, *et al.*, "Overview of recruitment for the osteoporotic fractures in men study (MrOS)," *Contemporary Clinical Trials*, vol. 26, no. 5, pp. 557–568, 2005. DOI: 10.1016/j.cct.2005.05.005.

[18] E. Orwoll, J. B. Blank, E. Barrett-Connor, *et al.*, "Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study — A large observational study of the determinants of fracture in older men," *Contemporary Clinical Trials*, vol. 26, no. 5, pp. 569–585, 2005. DOI: 10.1016/j.cct.2005.05.006.

[19] T. Blackwell, K. Yaffe, S. Ancoli-Israel, *et al.*, "Associations Between Sleep Architecture and Sleep-Disordered Breathing and Cognition in Older Community-Dwelling Men: The Osteoporotic Fractures in Men Sleep Study," *Journal of the American Geriatrics Society*, vol. 59, no. 12, pp. 2217–2225, 2011. DOI: 10.1111/j.1532-5415.2011.03731.x.

[20] D. A. Dean, A. L. Goldberger, R. Mueller, *et al.*, "Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource," *Sleep*, vol. 39, no. 5, pp. 1151–1164, 2016. DOI: 10.5665/sleep.5774.

[21] G.-Q. Zhang, L. Cui, R. Mueller, *et al.*, "The National Sleep Research Resource: towards a sleep data commons," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1351–1358, 2018. DOI: 10.1093/jamia/ocy064.

[22] S. Chambon, V. Thorey, P. J. Arnal, E. Mignot, and A. Gramfort, "A Deep Learning Architecture to Detect Events in EEG Signals During Sleep," in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, Aalborg, Denmark: IEEE, 2018, pp. 1–6. DOI: 10.1109/MLSP.2018.8517067.

[23] S. Chambon, V. Thorey, P. Arnal, E. Mignot, and A. Gramfort, "DOSED: A deep learning approach to detect multiple sleep micro-events in EEG signal," *Journal of Neuroscience Methods*, vol. 321, pp. 64–78, 2019. DOI: 10.1016/j.jneumeth.2019.03.017.

[24] A. N. Olesen, S. Chambon, V. Thorey, P. Jennum, E. Mignot, and H. B. D. Sorensen, "Towards a Flexible Deep Learning Method for Automatic Detection of Clinically Relevant Multi-Modal Events in the Polysomnogram," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany: IEEE, 2019, pp. 556–561. DOI: 10.1109/EMBC.2019.8856570.

[25] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches," arXiv: 1409.1259 [cs.CL].

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020. DOI: 10.1109/TPAMI.2018.2858826.

[27] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, pp. 1–15, 2015. arXiv: 1412.6980 [cs.LG].

[28] A. Paszke, S. Gross, F. Massa, *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advanced in Neural Information Processing Systems 32*, 2019, pp. 8024–8035. arXiv: 1912.01703 [cs.LG].

F

PAPER VI

TITLE: A multi-modal sleep event detectionmodel for clinical sleep analysis

AUTHORS: Alexander Neergaard Olesen, Poul Jennum, Emmanuel Mignot, and Helge B. D. Sorensen

STATUS: Manuscript in preparation

MSED: a multi-modal sleep event detection model for clinical sleep analysis

Alexander Neergaard Olesen, *Member, IEEE*, Poul Jennum,

Emmanuel Mignot, and Helge B. D. Sorensen, *Senior Member, IEEE*

April 25, 2020

Abstract

Study objective: Clinical sleep analysis require manual analysis of sleep patterns for correct diagnosis of sleep disorders. Several studies show significant variability in scoring discrete sleep events. We wished to investigate, whether an automatic method could be used for detection of arousals (Ar), leg movements (LM) and sleep disordered breathing (SDB) events, and if the joint detection of these events performed better than having three separate models.

Methods: We designed a single deep neural network architecture to jointly detect sleep events in a polysomnogram. We trained the model on 1653 recordings of individuals, and tested the optimized model on 1000 separate recordings. The performance of the model was quantified by F1, precision, and recall scores, and by correlating index values to clinical values using Pearson's correlation coefficient.

Results: F1 scores for the optimized model was 0.70, 0.63, and 0.62 for Ar, LM, and SDB, respectively. The performance was higher, when detecting events jointly compared to corresponding single-event models. Index values computed from detected events correlated well with manual annotations ($r^2 = 0.73$, $r^2 = 0.77$, $r^2 = 0.78$, respectively).

Conclusion: Detecting arousals, leg movements and sleep disordered breathing events jointly is possible, and the computed index values correlates well with human annotations.

1 Introduction

Clinical sleep analysis is currently performed manually by experts based on guidelines from the American Academy of Sleep Medicine (AASM) detailed in the AASM Scoring Manual [1]. The guidelines detail both technical and clinical best practices for setting up and recording polysomnographies (PSGs), which are overnight recordings of various electrophysiological signals, such as electroencephalography (EEG), electrooculography (EOG), chin and leg electromyography (EMG), electrocardiography (ECG), respiratory inductance plethysmography from the thorax and abdomen, oronasal pressure, and blood oxygen levels.

Based on these signals, expert technicians analyse and score the PSG for sleep stages [wakefulness (W), rapid eye movement (REM) sleep, non-REM stage 1 (N1), non-REM stage 2 (N2), and non-REM stage 3 (N3)], and sleep micro-events summarized in key metrics, such as the apnea-hypopnea index (AHI) (number of apneas and hypopneas per hour of sleep), the periodic leg movement index (PLMI) (number of period leg movements per hour of sleep), and the arousal index (ArI) (number of arousals per hour of sleep).

Arousals are defined as abrupt shifts in EEG frequencies towards alpha, theta, and beta rhythms for at least 3 s with a preceding period of stable sleep of at least 10 s. During REM sleep, where the background EEG shows similar rhythms, arousal scoring requires a concurrent increase in chin EMG lasting at least 1 s. Limb movements (LMs) should be scored in the leg EMG channels, when there is an increase in amplitude of at least 8 µV above baseline level with a duration between 0.5 s to 10 s. A periodic leg movement (PLM) series is then defined as a sequence of 4 LMs, where the time between LM onsets is between 5 min to 90 min. Apneas are generally scored when there is a complete (≥90 % of pre-event baseline) cessation of breathing activity either due to a physical obstruction (obstructive apnea) or due to an underlying disruption in the central nervous system control (central apnea) for at least 10 s. When the breathing is only partially reduced (≥30 % of pre-event baseline) and the duration of the excursion is ≥10 s, the event is scored as a hypopnea if there is either a ≥4 % oxygen desaturation or a ≥3 % oxygen desaturation coupled with an arousal (Ar).

However, several studies have shown significant variability in the scoring of both sleep stages [2]–[8] and sleep micro-events [9]–[16]. This has prompted extensive research into automatic methods for classifying sleep stages in large-scale studies [17]–[24], while the research in automatic arousal [25]–[27] and LM [28] detection on a similar scale is limited.

In this study, we introduce the multi-modal sleep event detection (MSED) model for joint detection of sleep micro-events, in this case Ars, sleep disordered breathing (SDB), and LM. The model is based on recent advances in machine learning and challenges current state of the art methods by directly classifying and localizing sleep micro-events in the PSG signals at the same time.

2 Data

We collected PSGs from the MrOS Sleep Study, an ancillary part of the larger Osteoporotic Fractures in Men Study. The main goal of the study is to research and discover connections between sleep disorders,

Table 1: MrOS demographics by subset.

	$\mathcal{D}_{\text{TRAIN}}$	$\mathcal{D}_{\text{EVAL}}$	$\mathcal{D}_{\text{TEST}}$	p-value
n	1653	200	1000	-
Age, years	76.4 ± 5.6 $[67.0, 90.0]$	76.8 ± 5.4 $[68.0, 90.0]$	76.4 ± 5.3 $[67.0, 90.0]$	0.404
BMI, kg s^{-2}	27.3 ± 3.9 $[16.0, 47.0]$	27.0 ± 3.6 $[19.0, 40.0]$	27.0 ± 3.7 $[17.0, 45.0]$	0.247
TST, min	357.3 ± 69.0 $[54.0, 615.0]$	354.0 ± 69.1 $[108.0, 503.0]$	353.6 ± 68.7 $[62.0, 572.0]$	0.312
SL, min	22.9 ± 25.6 $[1.0, 349.0]$	21.6 ± 23.0 $[1.0, 135.0]$	25.1 ± 32.1 $[1.0, 402.0]$	0.284
REML, min	109.5 ± 77.9 $[0.0, 578.0]$	103.5 ± 70.0 $[10.0, 413.0]$	107.2 ± 75.3 $[3.0, 590.0]$	0.466
WASO, min	116.7 ± 67.1 $[11.0, 462.0]$	119.0 ± 70.8 $[15.0, 372.0]$	112.9 ± 65.0 $[6.0, 458.0]$	0.471
SE, %	75.9 ± 12.1 $[17.0, 97.0]$	75.5 ± 12.3 $[37.0, 96.0]$	76.4 ± 11.8 $[26.0, 98.0]$	0.690
N1, %	6.8 ± 4.1 $[0.0, 31.0]$	7.0 ± 4.5 $[0.0, 28.0]$	6.9 ± 4.7 $[1.0, 58.0]$	0.968
N2, %	62.7 ± 9.5 $[28.0, 89.0]$	62.0 ± 9.7 $[30.0, 90.0]$	62.8 ± 10.0 $[21.0, 95.0]$	0.451
N3, %	11.4 ± 9.0 $[0.0, 55.0]$	11.8 ± 9.7 $[0.0, 55.0]$	11.1 ± 9.0 $[0.0, 57.0]$	0.638
REM, %	19.2 ± 6.5 $[0.0, 44.0]$	19.4 ± 7.2 $[0.0, 41.0]$	19.3 ± 6.7 $[0.0, 42.0]$	0.894
ArI, h^{-1}	23.5 ± 11.8 $[3.0, 87.0]$	23.4 ± 11.0 $[4.0, 77.0]$	23.8 ± 11.8 $[4.0, 102.0]$	0.661
AHI, h^{-1}	13.5 ± 13.9 $[0.0, 83.0]$	13.6 ± 13.3 $[0.0, 59.0]$	14.2 ± 15.5 $[0.0, 89.0]$	0.907
PLMI, h^{-1}	35.4 ± 37.1 $[0.0, 233.0]$	36.6 ± 39.0 $[0.0, 178.0]$	36.0 ± 37.7 $[0.0, 175.0]$	0.993

Significant p-values at significance level $\alpha = 0.05$ are highlighted in bold. BMI: body-mass index; TST: total sleep time; SL: sleep latency; REML: REM sleep latency; WASO: wake after sleep onset; SE: sleep efficiency; N1: non-REM stage 1; N2: non-REM stage 2; N3: non-REM stage 3; REM: rapid eye movement; ArI: arousal index; AHI: apnea-hypopnea index; PLMI: periodic leg movement index.

skeletal fractures, and cardiovascular disease and mortality in community-dwelling older (>65 years) [29]–[31]. Of the original 5994 study participants, 3135 subjects were enrolled at one of six sites in the USA for a comprehensive sleep assessment, while 2909 of these underwent a full-night in-home PSG recording, The PSG studies were subsequently scored by certified sleep technicians. Sleep stages were scored into stages 1, 2, 3, 4 and REM, while stages 3 and 4 combined into slow wave sleep (SWS) according to R&K rules [32]. Ars were scored as abrupt increases in EEG frequencies lasting at least 3 s according to American Sleep Disorders Association (ASDA) rules [33]. Apneas were defined as complete or near complete cessation of airflow lasting more than 10 s with an associated 3 % or greater SaO$_2$ desaturation, while hypopneas were based on a clear reduction in breathing of more than 30 % deviation from baseline breathing lasting more than 10 s, and likewise assocated with a greater than 3 % SaO$_2$ desaturation. While the scoring criteria for scoring LMs are not explicitly available for the MrOS Sleep Study, the prevailing standard at the time of the study was to score LMs following an increase in leg EMG amplitude of more than 8 µV above resting baseline levels for at least 0.5 s, but shorter than 10 s [34].

2.1 Subset demographics and partitioning

We used a total of 2853 PSG studies downloaded from the National Sleep Research Resource (NSRR) [35], [36], which we partitioned into a training set ($\mathcal{D}_{\text{TRAIN}}$, $n_{\text{train}} = 1653$), a validation set ($\mathcal{D}_{\text{EVAL}}$, $n_{\text{eval}} = 200$), and a final testing set ($\mathcal{D}_{\text{TEST}}$, $n_{\text{test}} = 1000$). Key demographics and PSG-related variables for each subset are shown as mean ± standard deviation with range in parenthesis in Table 1.

2.2 Signal and events

For this study, we considered three PSG events: Ars, LMs, and SDB events, which includes all forms of apneas (obstructive and central) and hypopneas. These event types are each based on a specific set of electrophysiological channels from the PSG, and as such, we extracted left and right central EEG (C3 and C4), left and right EOG, left and right chin EMG, left and right leg EMG, nasal pressure, and respiratory inductance plethysmography from the thorax and abdomen. EEG and EOG channels were referenced to the contralateral mastoid process, while a chin EMG was synthesized by subtracting the right chin EMG from the left chin EMG.

Apart from the raw signal data, we also extracted onset time relative to the study start time and duration times for each event type in each PSG.

3 Methods

Notation We denote by $[\![a, b]\!]$ the set of integers $\{n \in \mathbb{N} \mid a \leq n \leq b\}$ with $[\![N]\!]$ being shorthand for $[\![1, N]\!]$, and by $n \in [\![N]\!]$ the nth sample in $[\![N]\!]$. A segment of PSG data is denoted by $\mathbf{x} \in \mathbb{R}^{C \times T}$, where C, T is the number of channels and the duration of the segment in samples, respectively. The corresponding set of N_t true events for the segment is denoted by $\boldsymbol{\varepsilon}^t = \left\{(\varrho_i^t, \delta_i^t) \in \mathbb{R}_+^2 \mid i \in [\![N_t]\!]\right\}$, where ϱ, δ are the center point and duration, respectively, of the ith event. By $\chi \in \mathcal{D}_*$ we denote a sample in either one of the three subsets. In the description of the network architecture, we have omitted the batch dimension from all calculations for brevity.

3.1 Model overview

Given an input set $\boldsymbol{\chi} = \{\mathbf{x}, \boldsymbol{\varepsilon}^t\} \in \mathbb{R}^{C \times T} \times \mathbb{R}_+^{N_t \times 2}$ containing PSG data with C channels and T time steps, and true events $\boldsymbol{\varepsilon}$, the goal of the model is to detect any possible events in the segment, where, in this context, detection covers both classification *and* localization of any event in the segment space.

To accomplish this, the model generates a set of *default event windows* $\boldsymbol{\varepsilon}^d = \left\{\left(\varrho_j^d, \delta_j^d\right) \in \mathbb{R}_+^2 \mid j \in [\![N_d]\!]\right\}$ for the current segment, and match each true event to a default event window if their intersection-over-union (IoU) is at least 0.5.

At test time, we generate predictions over the default event windows and use a non-maximum suppression procedure to select between the candidate predictions. For a given class k, the procedure is as follows. First, the predictions are sorted according to probability of the event, which is above a threshold θ_k. Then, using the most probable prediction as an anchor, we sequentially evaluate the IoU between the anchor and the remaining candidate predictions, removing those with IoU $>= 0.5$.

The output of the model is thus the set $\{\mathbf{p}, \mathbf{y}\}$ containing the predicted class probabilities along with the corresponding onsets and durations.

3.2 Signal conditioning

We resampled all signals to a common sampling frequency of $f_s = 128\,\text{Hz}$ using a poly-phase filtering approach (Kaiser window, $\beta = 5.0$). Based on recommended filter specifications from the AASM, we designed Butterworth IIR filters for four sets of signals. EEG and EOG channels were filtered with a 2nd order filter with a 0.3 Hz–35 Hz passband, while chin and leg EMG channels were filtered with a 4th order high-pass filter with a 10 Hz cut-off frequency. The nasal pressure channel was filtered with a 4th order high-pass filter with a 0.03 Hz cut-off frequency, while the thoracoabdominal channels were filtered with a 2nd order with a 0.1 Hz–15 Hz passband.

All filters were implemented using the zero-phase method, which sequentially applies the filter in the forward direction, and then in the backwards direction. This accounts for the non-linear phase response and subsequent frequency-dependent group delay inherent in IIR filters, but also effectively squares the magnitude response of the filter.

Filtered signals were subsequently standardized by

$$\mathbf{x}^{(i)} = \frac{\hat{\mathbf{x}}^{(i)} - \boldsymbol{\mu}^{(i)}}{\boldsymbol{\sigma}^{(i)}}, \tag{1}$$

where $\hat{\mathbf{x}}^{(i)} \in \mathbb{R}^{C \times T}$ is the raw matrix containing C input channels and T samples, and $\boldsymbol{\mu}^{(i)}, \boldsymbol{\sigma}^{(i)} \in \mathbb{R}^{C}$ are the mean and standard deviation vectors for the ith PSG, respectively. This is a common approach in computer vision tasks, and beneficial to ensure a proper gradient propagation through a deep neural network [37].

3.3 Target encoding

For each data segment, target event classes $\boldsymbol{\pi} \in \mathbb{R}^{N_m \times K}$ generated by one-hot encoding, while the target detection variable containing the onset and duration times $\mathbf{t} \in \mathbb{R}^{N_m \times 2}$ was encoded as

$$t_i = \left(\frac{\varrho_i^m - \varrho_j^d}{\delta_j^d}, \log \frac{\delta_i^m}{\delta_j^d} \right), \quad i \in [\![N_m]\!],\, j \in [\![N_d]\!], \tag{2}$$

where ϱ_i^m is the center point of the true event matched to a default event window ϱ_j^d, and δ_i^m and δ_j^d are the corresponding durations of the true and default events.

3.4 Data sampling

As the total number of default event windows in a data segment N_d most likely will be much higher than the number of event windows matched to a true event, i. e. $N_d \gg N_m$, we implemented a similar random data sampling strategy as in [25]. At training step t, a given PSG record r has a certain number of associated number of Ar, LM, and SDB events ($n_{\text{Ar}}, n_{\text{LM}}, n_{\text{SDB}}$, respectively). We randomly sample a class k with equal probability $p_k = \frac{1}{K}$, whilst disregarding the negative class, since this class is most likely over-represented in the data segment. Given the class k, we randomly sample an event ε_k with probability $p(\varepsilon_k) \propto n_k$, and afterwards, we extract a segment of data of size $C \times T$, where the start of the segment is sampled from $[\bar{\varepsilon}_k - T, \bar{\varepsilon}_k + T]$, where $\bar{\varepsilon}_k$ is the sample midpoint of the event ε_k, thereby ensuring at least 50 % overlap with at least one event associated with the data segment.

We found that this approach to sampling data segments with a large ratio of negative to positive samples to be beneficial in all our experiments, when monitoring the loss on the validation set.

3.5 Network architecture

Similar to the architecture described in [27], we designed a splitstream network architecture for the differentiable function Φ, where each stream is responsible for the bulk feature extraction for a specific event class. For the given problem of detecting Ars, LMs, and SDBs, the network contains three streams: the Ar stream takes as input the EEGs, the EOGs, and the chin EOG signals for a total of $C_{\mathrm{Ar}} = 5$ channels; the LM stream receives the $C_{\mathrm{LM}} = 2$ leg EMG signals; and the SDB stream receives the nasal pressure and the thoracoabdominal signals for a total of $C_{\mathrm{SDB}} = 3$ channels. An overview of the network architecture is shown graphically in Figure 1.

3.5.1 Stream specifics

Each stream is comprised of two components. First, a mixing module $\varphi_{\mathrm{mix}} : \mathbb{R}^{C_* \times T} \to \mathbb{R}^{C_* \times T}$ computes a non-linear mixing of the C channels using a set of C single-strided 1-dimensional filters $\mathbf{w} \in \mathbb{R}^{C \times C}$ and rectified linear unit (ReLU) activation [38], such that $\varphi_{\mathrm{mix}}(\mathbf{x}) = \max\{0, \mathbf{w} \otimes \mathbf{x} + \mathbf{b}\}$, where the max operation introduces the non-linearity, $\otimes$ is the convolution (conv) operator over the C feature maps, and $\mathbf{b} \in \mathbb{R}^{C}$ is a bias vector (in this case $\mathbf{b} = 0$). Second, the output activations from φ_{mix} are used as input to a deep neural network module $\varphi_{\mathrm{feat}} : \mathbb{R}^{C_* \times T} \to \mathbb{R}^{f' \times T'}$, which transforms the input feature maps to a $f' \times T'$ feature space with a temporal dimension reduced by a factor of $\frac{T}{T'}$. The feature extraction module φ_{feat} is realized using $k_{\max}$ successive conv operations with an increasing number of filters $f' = f_0 2^{k-1}$, $k \in [\![k_{\max}]\!]$, where f_0 is a tunable base filter number. Each conv feature map is normalized using batch normalization (BN) [39], such that if $\tilde{\mathbf{z}} \in \mathbb{R}^{f' \times T'}$ denotes the output from a conv operation, the subsequent normalized version is computed as

$$\mathbf{z} = \gamma \frac{\tilde{\mathbf{z}} - \mathrm{E}[\tilde{\mathbf{z}}]}{\sqrt{\mathrm{Var}[\tilde{\mathbf{z}}] + \epsilon}} + \boldsymbol{\beta}, \tag{3}$$

where $\mathrm{E}[\tilde{\mathbf{z}}] \in \mathbb{R}^{f'}$, $\mathrm{Var}[\tilde{\mathbf{z}}] \in \mathbb{R}_{+}^{f'}$ is the expectation and variance over the temporal dimension of each feature map, ϵ is a small constant, and $\{\boldsymbol{\gamma}, \boldsymbol{\beta}\} \in \mathbb{R}^{f'} \times \mathbb{R}^{f'}$ are learnable parameters representing the mean and bias for each feature map. Each normalized conv output is subsequently activated using ReLU.

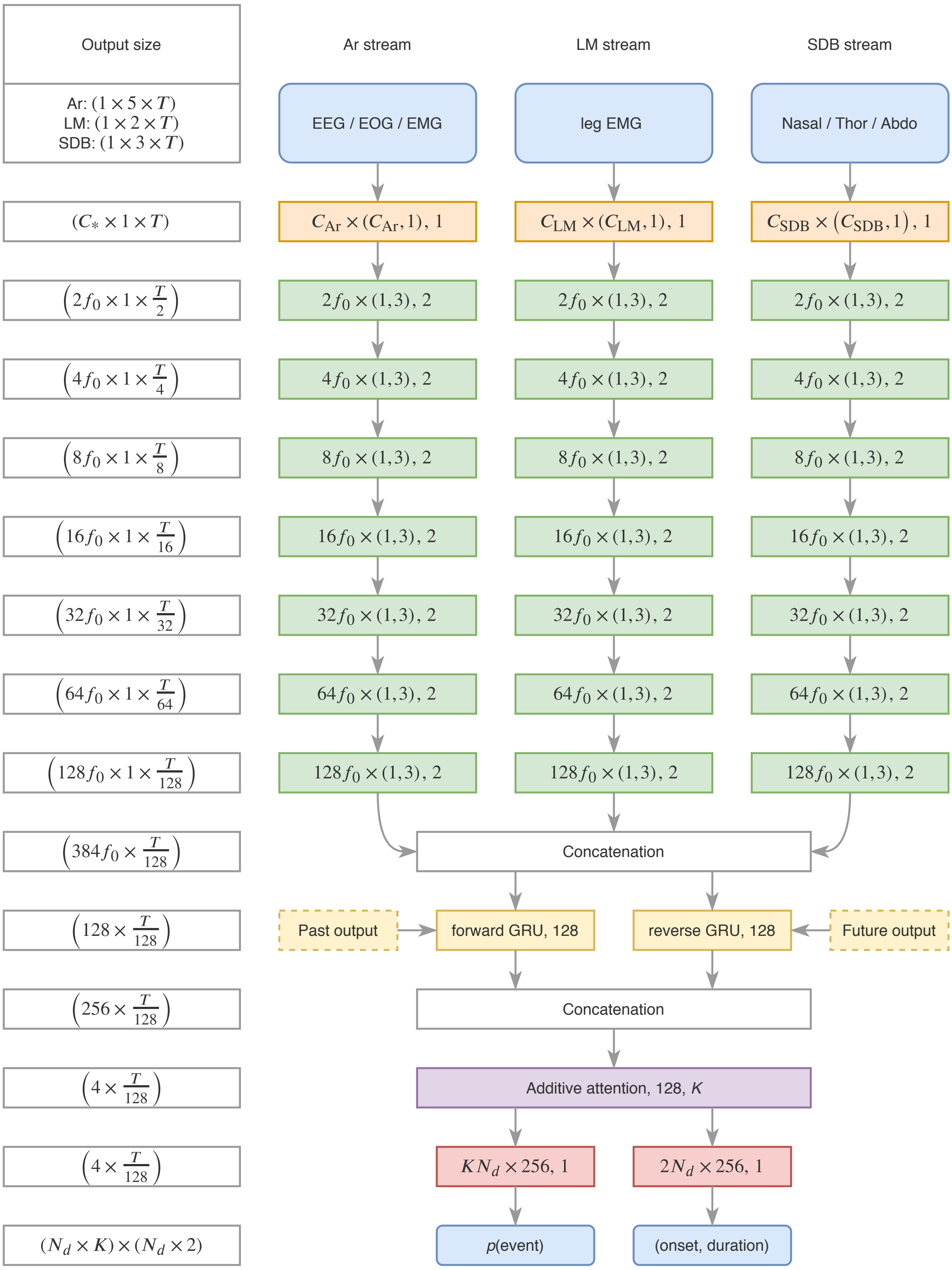

Output size
Ar: $(1 \times 5 \times T)$
LM: $(1 \times 2 \times T)$
SDB: $(1 \times 3 \times T)$
Ar stream
LM stream
SDB stream
EEG / EOG / EMG
leg EMG
Nasal / Thor / Abdo
$(C_{*} \times 1 \times T)$
$C_{\text{Ar}} \times (C_{\text{Ar}}, 1)$, 1
$C_{\text{LM}} \times (C_{\text{LM}}, 1)$, 1
$C_{\text{SDB}} \times (C_{\text{SDB}}, 1)$, 1
$\left(2f_0 \times 1 \times \frac{T}{2}\right)$
$2f_0 \times (1,3)$, 2
$\left(4f_0 \times 1 \times \frac{T}{4}\right)$
$4f_0 \times (1,3)$, 2
$\left(8f_0 \times 1 \times \frac{T}{8}\right)$
$8f_0 \times (1,3)$, 2
$\left(16f_0 \times 1 \times \frac{T}{16}\right)$
$16f_0 \times (1,3)$, 2
$\left(32f_0 \times 1 \times \frac{T}{32}\right)$
$32f_0 \times (1,3)$, 2
$\left(64f_0 \times 1 \times \frac{T}{64}\right)$
$64f_0 \times (1,3)$, 2
$\left(128f_0 \times 1 \times \frac{T}{128}\right)$
$128f_0 \times (1,3)$, 2
$\left(384f_0 \times \frac{T}{128}\right)$
Concatenation
$\left(128 \times \frac{T}{128}\right)$
Past output
forward GRU, 128
reverse GRU, 128
Future output
$\left(256 \times \frac{T}{128}\right)$
Concatenation
$\left(4 \times \frac{T}{128}\right)$
Additive attention, 128, K
$\left(4 \times \frac{T}{128}\right)$
$KN_d \times 256$, 1
$2N_d \times 256$, 1
$(N_d \times K) \times (N_d \times 2)$
p(event)
(onset, duration)

Figure 1: MSED network architecture. The left column shows the output dimensions for each operation as (number of filters[x singleton] x time steps). Each stream on the right (green) processes a separate set of input channels (blue, top), the results of which are concatenated before the bidirectional gated recurrent unit (bGRU) (yellow). The outputs from the additive attention layer (purple) are convolved in the final classification and localization layers (red) to output the probabilities for each event class, and the predicted onset and duration of each event (blue, bottom). Convolution layers (orange, green, red) are detailed as [number of feature maps x kernel size, stride]. Recurrent layer (yellow) shows the direction and number of hidden units. Additive attention layer (purple) is described with the number of hidden and output units.

3.5.2 Feature fusion for sequential processing

The outputs from the three feature extraction streams are subsequently fused by concenating each output vector $\mathbf{z}_*$ into a combined feature vector $\mathbf{z} = (\mathbf{z}_{\text{ar}}, \mathbf{z}_{\text{lm}}, \mathbf{z}_{\text{sdb}}) \in \mathbb{R}^{3f' \times T'}$. We introduce sequential modeling of the feature vectors using a bGRU [40], which has the advantage over other recurrent neural network (RNN)-based models such as the long short-term memory (LSTM) of having fewer trainable parameters while still being powerful enough to model complex, temporal relationships [41]. The output of the gated recurrent unit (GRU) for timestep t is a vector $\mathbf{h}_t = \left(\mathbf{h}_t^{\mathsf{f}}, \mathbf{h}_t^{\mathsf{b}}\right) \in \mathbb{R}^{2n_h}$ containing the concatenated outputs from the forward (f) and backward (b) directions. Each directional feature vector is calculated as a weighted combination of a gated new input $\mathbf{n}_t$ and the feature vector from the previous timestep $\mathbf{h}_{t-1}$

$$\mathbf{h}_t^* = (1 - \mathbf{u}_t) \otimes \mathbf{n}_t + \mathbf{u}_t \otimes \mathbf{h}_{t-1}. \tag{4}$$

The update gate $\mathbf{u}_t$ and gated new input $\mathbf{n}_t$ are computed as

$$\mathbf{u}_t = \sigma\big(\mathbf{W}_u^z \mathbf{z}_t + \mathbf{b}_u^z + \mathbf{W}_u^h \mathbf{h}_{t-1} + \mathbf{b}_u^h\big), \tag{5}$$

$$\mathbf{n}_t = \tanh\Big(\mathbf{W}_n^z \mathbf{z}_t + \mathbf{b}_n^z + \mathbf{r}_t \otimes \big(\mathbf{W}_n^h \mathbf{h}_{t-1} + \mathbf{b}_n^h\big)\Big), \tag{6}$$

where $\mathbf{W}_*^*, \mathbf{b}_*^*$ are weight matrices and bias vectors, respectively, and $\mathbf{r}_t$ is a reset gate computed as

$$\mathbf{r}_t = \sigma(\mathbf{W}_r^z \mathbf{z}_t + \mathbf{b}_r^z + \mathbf{W}_h^r \mathbf{h}_{t-1} + \mathbf{b}_h^r). \tag{7}$$

3.5.3 Additive attention

The attention mechanism is a powerful technique to introduce a way for the network to focus on relevant regions and disregard irrelevant regions of a data sample, and is a key part of the highly successful Transformer model [42] and the subsequent state-of-the-art BERT model for natural language processing [43]. In this work, we implemented a simple, but powerful, *additive attention* mechanism [44], which computes *context*-vectors $\mathbf{c} \in \mathbb{R}^{2n_h}$ for each event class as the weighted sum of the feature vector outputs $\mathbf{h} \in \mathbb{R}^{2n_h \times T'}$ from the φ_h. Formally, attention is computed as

$$\mathbf{c} = \mathbf{h} \cdot \boldsymbol{\alpha} = \sum_{t=1}^{T'} \mathbf{h}_t \boldsymbol{\alpha}_t, \tag{8}$$

where T' is the reduced temporal dimension, $\mathbf{h}_t$ is the feature vector for time step t, and $\boldsymbol{\alpha}_t \in \mathbb{R}^K$ is the attention weight computed as

$$\boldsymbol{\alpha}_t = \frac{\exp(\tanh(\mathbf{h}_t \mathbf{W}_{\text{u}}) \mathbf{W}_a)}{\sum_{\tau}^{T'} \exp(\tanh(\mathbf{h}_\tau \mathbf{W}_{\text{u}}) \mathbf{W}_a)}. \tag{9}$$

Here, $\mathbf{W}_u \in \mathbb{R}^{2n_h \times n_a}$ and $\mathbf{W}_a \in \mathbb{R}^{n_a \times K}$ are linear mappings of the feature vectors, and tanh is the hyperbolic tangent function.

3.5.4 Detection

The final event classification and localization is handled by two modules, $\psi_{\text{clf}} : \mathbb{R}^{2n_h \times K} \to \mathbb{R}^{N_d \times K}$ and $\psi_{\text{loc}} : \mathbb{R}^{2n_h \times K} \to \mathbb{R}^{N_d \times 2}$, respectively. The classification module $\psi_{\text{clf}} : \mathbf{c} \mapsto \mathbf{p}$ outputs a tensor $\mathbf{p} \in [0,1]_+^{N_d \times K}$ containing predicted event class probabilities for each default event window. The localization module $\psi_{\text{loc}} : \mathbf{c} \mapsto \mathbf{y}$ outputs a tensor $\mathbf{y} \in \mathbb{R}^{N_d \times 2}$ containing encoded relative onsets and durations for a detected event for each default event window.

3.6 Loss function

Similar to [45], we optimized the network parameters according to a three-component loss function consisting of: i) a localization loss ℓ_{loc}; ii) a positive classification loss ℓ_+, and iii) a negative classification loss ℓ_-, such that the total loss ℓ was defined by

$$\ell = \ell_{\text{loc}} + \ell_+ + \ell_-. \tag{10}$$

The localization loss ℓ_{loc} was calculated using a Huber function

$$\ell_{\text{loc}} = \frac{1}{N_+} \sum_{i \in \pi_+} f_H^{(i)} \tag{11}$$

$$\mathbf{f}_H = \begin{cases} 0.5(\mathbf{y} - \mathbf{t})^2, & \text{if } |\mathbf{y} - \mathbf{t}| < 1, \\ |\mathbf{y} - \mathbf{t}| - 0.5, & \text{otherwise,} \end{cases} \tag{12}$$

where $i \in \pi_+$ yields indices of event windows with positive targets, i. e. event windows matched to an arousal, LM or SDB target, and N_+ is the number of positive targets in the given data segment.

The positive classification loss component ℓ_+ was calculated using a simple cross-entropy over the event windows matched to an arousal, LM, or SDB event:

$$\ell_+ = \frac{1}{N_+} \sum_{i \in \pi_+} \sum_{k \in [\![K]\!]} \pi_k^{(i)} \log p_k^{(i)}, \quad \text{where} \quad p_k^{(i)} = \frac{\exp s_k^{(i)}}{\sum_j \exp s_j^{(i)}}, \tag{13}$$

and $\pi_k^{(i)}$, $p_k^{(i)}$, and $s_k^{(i)}$ are the true class probability, predicted class probability, and logit score for the ith event window containing a positive sample.

Similar to [46], [47], the negative classification loss ℓ_- was calculated using a hard negative mining approach to balance the number of positive and negative samples in a data segment after matching default event windows to true events [48]. Specifically, this is accomplished by calculating the probability for the negative class (no event) for each unmatched default event window, and then calculating the cross entropy loss using the Z most probable samples. In our experiments, we set the ratio of positive to negative samples as 1:3, such that the calculation of ℓ involves $Z = 3$ times as many negative as positive samples.

We also explored a focal loss objective function for computing ℓ_+ and ℓ_- [49], however, we found that this approach severely deteriorated the ability of the network to accurately detect LM and SDB events compared to using worst negative mining.

3.7 Optimization

The network parameters were optimized using adaptive moment estimation (Adam) according to the loss function described in Equation (10) [50]. This algorithms uses first (m) and second (v) moment estimations of gradients to update the model parameters θ of a differentiable function f at time t:

$$m^{(t)} = \beta_1 m^{(t-1)} + (1 - \beta_1)\nabla_\theta f^{(t)}\left(\theta^{(t-1)}\right) \tag{14}$$

$$v^{(t)} = \beta_2 v^{(t-1)} + (1 - \beta_2)\nabla_\theta^2 f^{(t)}\left(\theta^{(t-1)}\right), \tag{15}$$

where β_1, β_2 are exponential decay rates for the first and second moment, respectively, ∇ is the gradient vector with respect to θ, and ∇_θ^2 is the Hadamard product $\nabla_\theta f \odot \nabla_\theta f$. The moment vectors are initialized with 0's, which induce a bias towards zero. This can be offset by computing a bias-corrected estimate of each moment vector as

$$\hat{m}^{(t)} = \frac{m^{(t)}}{1 - \beta_1^t} \tag{16}$$

$$\hat{v}^{(t)} = \frac{v^{(t)}}{1 - \beta_2^t}, \tag{17}$$

which yields the final update to θ as

$$\theta^{(t)} = \theta^{(t-1)} - \eta \frac{\hat{m}^{(t)}}{\sqrt{\hat{v}^{(t)}} + \epsilon}, \tag{18}$$

where η is the learning rate.

3.8 Experimental setups

In our experiments, we fixed the exponential decay rates at $(\beta_1, \beta_2) = (0.9, 0.999)$, the learning rate at $\eta = 10^{-3}$, and $\epsilon = 10^{-8}$. The learning rate was decayed in a step-wise manner by multiplying η with a factor of 0.1 after 3 consecutive epochs with no improvement in loss value on the validation dataset.

Similarly, we employed an early stopping scheme by monitoring the loss on the validation dataset and

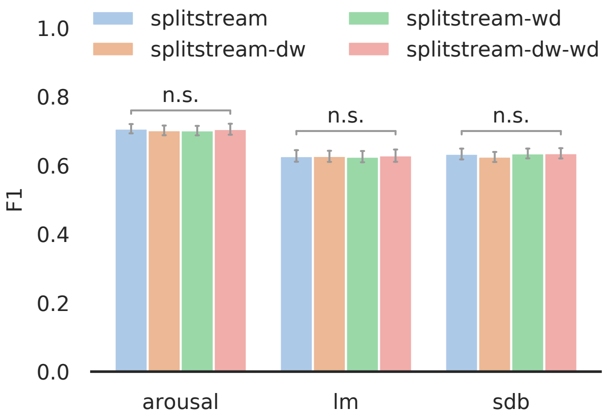


Figure 2: Caption

stopping the model training after 10 epochs of no improvement on $\mathcal{D}_{\text{EVAL}}$.

We tested four types of models in two categories: the first is a default split-stream model as shown in Figure 1 with and without weight decay (splitstream, splitstream-wd). The second is a variation of the split-stream model, but where the ψ_{clf} and ψ_{loc} modules are realized using depth-wise convolutions, such that each attention group is used only for that type of event. The second category is also tested with and without weight decay (splitstream-dw, splitstream-dw-wd).

3.9 Performance evaluation

Performance was quantified using precision, recall and F1 scores. Statistical significance in F1 score between groups was assessed with Kruskall-Wallis H-tests. The performance of joint vs. single-event detection models was tested with Wilcoxon signed rank tests for matched samples. The relationships between true and predicted ArI, AHI, and limb movement index (LMI) were assessed using linear models and Pearsons r^2. Significance was set at $\alpha = 0.05$.

3.10 Statistical analysis

We used Kruskall-Wallis H-tests for significant differences in demographic and PSG variables between subsets, and to test for significant differences in F1 performance between model architectures evaluated on $\mathcal{D}_{\text{EVAL}}$. Significant differences between joint and single-event detection models were assessed using Wilcoxon signed-rank tests. The level of significance for all tests were set at $\alpha = 0.05$.

4 Results and discussion

4.1 Model architecture evaluation

We found no significant differences in F1 performance for either Ar (Kruskal-Wallis $H = 0.961$, $p = 0.811$), LM ($H = 0.230$, $p = 0.973$), or SDB detection ($H = 2.838$, $p = 0.417$), when evaluating the model

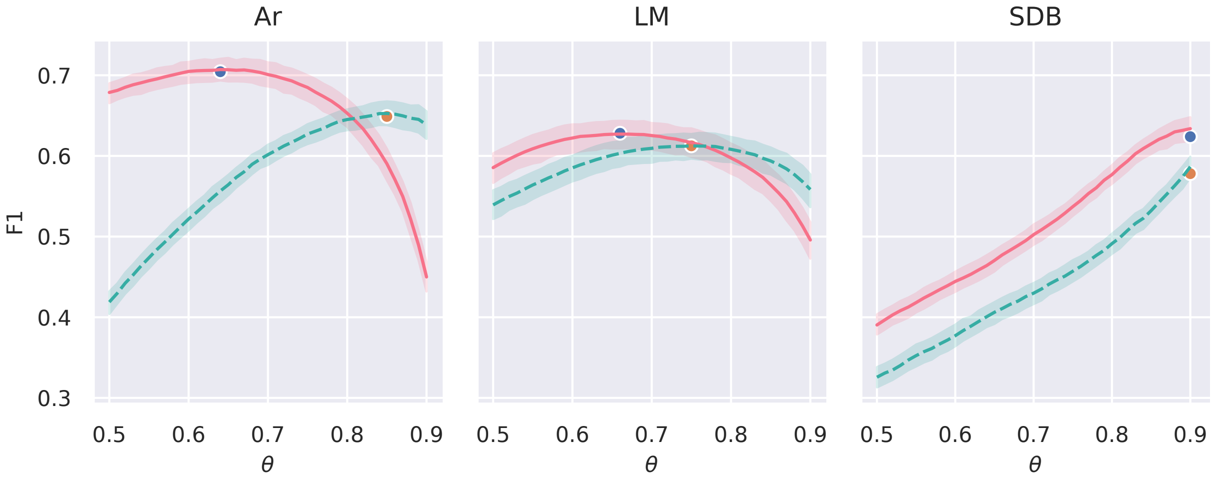


Figure 3: Optimizing F1 performance on $\mathcal{D}_{\text{EVAL}}$ as a function of θ). Full lines correspond to the joint model and dashed lines are the corresponding single-event detection model. The blue and orange dots correspond to optimized model performance on $\mathcal{D}_{\text{TEST}}$.

Table 2: Performance scores for optimized models evaluated on $\mathcal{D}_{\text{TEST}}$.

Event	**Model**	**Precision**	**Recall**	**F1**
Ar	Joint	0.759 ± 0.114	0.672 ± 0.125	0.704 ± 0.106
	Single	0.777 ± 0.107	0.571 ± 0.127	0.649 ± 0.113
LM	Joint	0.650 ± 0.169	0.647 ± 0.120	0.628 ± 0.123
	Single	0.661 ± 0.166	0.607 ± 0.116	0.613 ± 0.116
SDB	Joint	0.817 ± 0.142	0.526 ± 0.146	0.624 ± 0.115
	Single	0.765 ± 0.142	0.486 ± 0.121	0.578 ± 0.097

architectures on $\mathcal{D}_{\text{EVAL}}$. Based on this result, all further modeling was based on the default splitstream architecture for simplicity.

4.2 Joint vs. single event detection

For each event type, we evaluated the F1 score as a function of classification threshold θ on $\mathcal{D}_{\text{EVAL}}$ for both the joint detection model as well as the single-event models. It can be observed in Figure 3 that for all three events, the joint detection model achieves higher F1 score, although the apparent increase is not as large for LM detection. This was also observed when evaluating the joint and single detection models with optimized thresholds on $\mathcal{D}_{\text{TEST}}$ for both Ar (Wilcoxon $W = 30440.0$, $p = 2.481 \times 10^{-127}$), LM ($W = 101103.0$, $p = 6.454 \times 10^{-60}$), and SDB detection ($W = 93647.0$, $p = 2.378 \times 10^{-64}$). Precision, recall and F1 scores for optimized models evaluated on $\mathcal{D}_{\text{TEST}}$ are shown in Table 2. These findings are interesting, because they provide evidence that the presence of different event types can module the detection of others, and that this can be modeled using automatic methods. This is in line with what previous studies have found e. g. on event-by-event scoring agreement in arousals, which improved significantly from 0.59 % to 0.91 %, when including respiratory signals in the analysis [13].

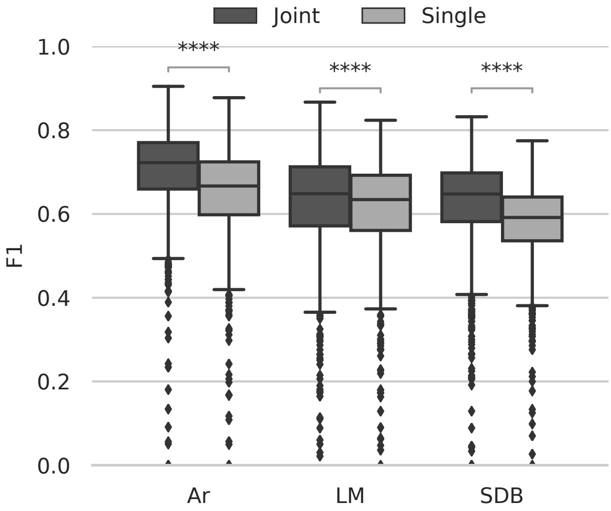


Figure 4: Evaluating optimized joint and single-event detection models on $\mathcal{D}_{\text{TEST}}$. ****: $p < 10 \times 10^{-4}$. Ar: arousal; LM: limb movement; SDB: sleep disordered breathing.

4.3 Detection vs. manual scorings

For each event type, we computed the correlation coefficient between the predicted and true index values (arousal index, ArI; apnea-hypopnea index, AHI; limb movement index, LMI), which is shown in Figure 5. We found a large positive correlation between true and predicted values for ArI ($r^2 = 0.73$, $p = 2.5 \times 10^{-285}$), AHI ($r^2 = 0.77$, $p = 9.3 \times 10^{-316}$), and LMI ($r^2 = 0.78$, $p = 3.1 \times 10^{-321}$).

A similar study using an automatic method for automatic detection of SDB[1] and LM events found similar or higher correlations between automatic and manual scorings ($r^2 = 0.85$, and $r^2 = 0.79$, respectively), although their findings were based on almost 5 times as much data [21].

4.3.1 Temporal characteristics

We compared the temporal precision between manual and automatic event scoring by looking at the errors in onset (Δonset), offsets (Δoffset), and durations (Δdur.) calculated as

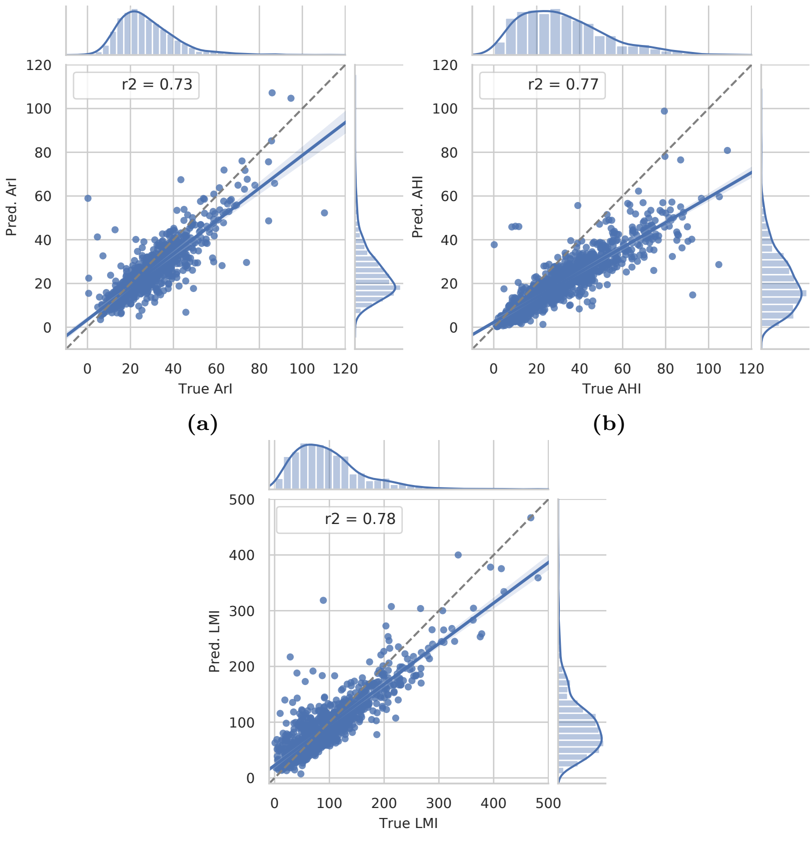
$$\Delta\,\text{onset} = \text{onset}_{\text{automatic}} - \text{onset}_{\text{manual}} \tag{19}$$

$$\Delta\,\text{offset} = \text{offset}_{\text{automatic}} - \text{offset}_{\text{manual}} \tag{20}$$

$$\Delta\,\text{dur} = \text{dur}_{\text{automatic}} - \text{dur}_{\text{manual}} \tag{21}$$

so that positive values of $\Delta\,\text{onset}, \Delta\,\text{offset}$ corresponds to a positive shift to the right (delayed prediction), and positive values of $\Delta\,\text{dur}$. meaning an overestimation of the event duration compared to manual scoring. This is shown in Figure 6, where the blue distributions are the joint detection model for each event type, and the orange distributions are the corresponding single-event models. The distributions are shown as kernel density estimates superimposed on a histogram. For Ar events, the model overestimates

[1] The authors pooled obstructive, central, mixed apneas, and 4% hypopneas into one category, *apnea*.



(a)

(b)

(c)

Figure 5: Pearson correlation plots for each event type index between true and predicted values. The linear relationship is indicated with solid blue with 95% confidence intervals in light blue. Grey dashed lines indicate perfect correlation lines. ArI: arousal index; AHI: apnea-hypopnea index; LMI: limb movement index.

the duration on average by a couple of seconds, which is caused by an earlier prediction of onset and delayed prediction of termination. For LM events, the model underestimates the duration by about half a second on average, which is due to earlier prediction of termination. For SDB events, the model overestimates the duration by about 25 seconds on average, which is caused by an earlier prediction of onset and delayed prediction of termination. These errors in predicted durations reflects the temporal characteristics of these events; LMs are shorter events[2], and it is thus unlikely to be overestimated by several seconds, while SDBs are longer events by one to two orders of magnitude, which also increases the size of the errors. Ars events are intermediate in length compared to LMs and SDBs, which is reflected in the error distributions.

References

[1] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. L. Marcus, and B. V. Vaughn, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.6.* Darien, IL, USA: American Academy of Sleep Medicine, 2020.

[2] Between 0.5 s to 10 s per definition.

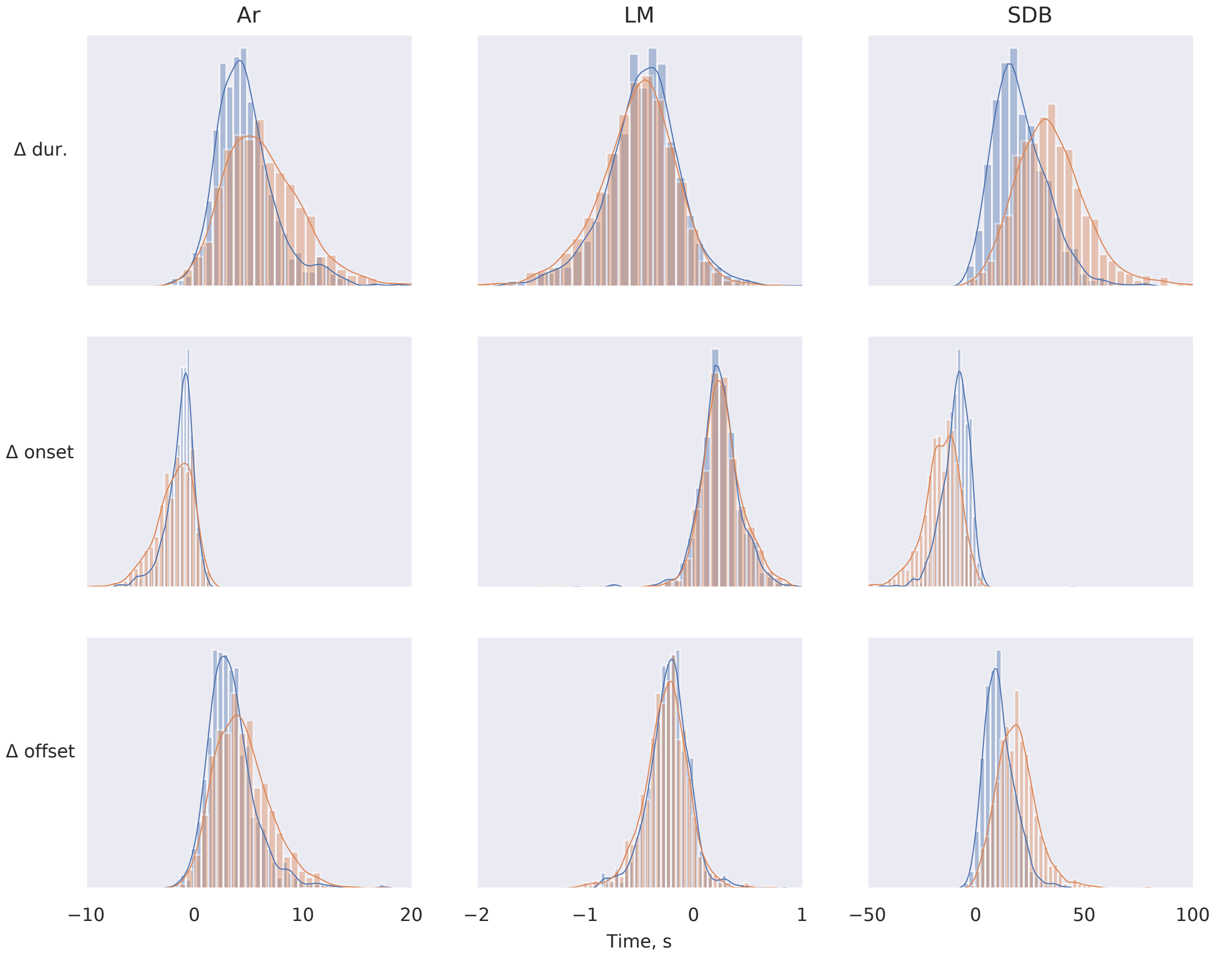


Figure 6: Temporal error metrics distributions across all events and PSGs. Positive values of Δonset, Δoffset means delayed predictions, while positive values of Δdur. means to an overestimation of event duration. Blue distributions are joint detection models, while orange distributions are the corresponding single-event models. Distributions are shown as kernel density estimates superimposed on a histogram. Ar: arousal; LM: limb movement; SDB: sleep disordered breathing.

[2] R. G. Norman, I. Pal, C. Stewart, J. A. Walsleben, and D. M. Rapoport, "Interobserver Agreement Among Sleep Scorers From Different Centers in a Large Dataset," *Sleep*, vol. 23, no. 7, pp. 1–8, 2000. DOI: 10.1093/sleep/23.7.1e.

[3] H. Danker-Hopfe, D. Kunz, G. Gruber, G. Klösch, J. L. Lorenzo, S. L. Himanen, B. Kemp, T. Penzel, J. Röschke, H. Dorn, A. Schlögl, E. Trenker, and G. Dorffner, "Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders," *J. Sleep Res.*, vol. 13, pp. 63–69, 2004. DOI: 10.1046/j.1365-2869.2003.00375.x.

[4] 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," *J. Sleep Res.*, vol. 18, no. 1, pp. 74–84, 2009. DOI: 10.1111/j.1365-2869.2008.00700.x.

[5] R. S. Rosenberg and S. Van Hout, "The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring," *J. Clin. Sleep Med.*, vol. 9, pp. 81–87, 2013. DOI: 10.5664/jcsm.2350.

[6] 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 Breath.*, vol. 19, no. 1, pp. 191–195, 2015. DOI: 10.1007/s11325-014-0990-0.

[7] M. Younes, J. Raneri, and P. Hanly, "Staging sleep in polysomnograms: Analysis of inter-scorer variability," *J. Clin. Sleep Med.*, vol. 12, no. 6, pp. 885–894, 2016. DOI: 10.5664/jcsm.5894.

[8] M. Younes, S. T. Kuna, A. I. Pack, J. K. Walsh, C. A. Kushida, B. Staley, and G. W. Pien, "Reliability of the American Academy of Sleep Medicine Rules for Assessing Sleep Depth in Clinical Practice," *J. Clin. Sleep Med.*, vol. 14, no. 2, pp. 205–213, 2018. DOI: 10.5664/jcsm.6934.

[9] M. J. Drinnan, A. Murray, C. J. Griffiths, and G. J. Gibson, "Interobserver Variability in Recognizing Arousal in Respiratory Sleep Disorders," *Am. J. Respir. Crit. Care Med.*, vol. 158, pp. 358–362, 1998. DOI: 10.1164/ajrccm.158.2.9705035.

[10] C. W. Whitney, D. J. Gottlieb, S. Redline, R. G. Norman, R. R. Dodge, E. Shahar, S. Surovec, and F. J. Nieto, "Reliability of scoring respiratory disturbance indices and sleep staging," *Sleep*, vol. 21, no. 7, pp. 749–757, 1998. DOI: 10.1093/sleep/21.7.749.

[11] J. S. Loredo, J. L. Clausen, S. Ancoli-Israel, and J. E. Dimsdale, "Night-to-Night Arousal Variability and Interscorer Reliability of Arousal Measurements," *Sleep*, vol. 22, no. 7, pp. 916–920, 1999. DOI: 10.1093/sleep/22.7.916.

[12] M. Smurra, M. Dury, G. Aubert, D. Rodenstein, and G. Liistro, "Sleep fragmentation: comparison of two definitions of short arousals during sleep in OSAS patients," *Eur. Respir. J.*, vol. 17, pp. 723–727, 2001. DOI: 10.1183/09031936.01.17407230.

[13] R. J. Thomas, "Arousals in Sleep-disordered Breathing: Patterns and Implications," *Sleep*, vol. 26, no. 8, pp. 1042–1047, 2003. DOI: 10.1093/sleep/26.8.1042.

[14] M. H. Bonnet, K. Doghramji, T. Roehrs, E. J. Stepanski, S. H. Sheldon, A. S. Walters, M. Wise, and A. L. Chesson, "The scoring of arousal in sleep: Reliability, validity, and alternatives," *J. Clin. Sleep Med.*, vol. 3, no. 2, pp. 133–145, 2007. DOI: 10.5664/jcsm.26815.

[15] 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*, vol. 36, no. 4, pp. 591–596, 2013. DOI: 10.5665/sleep.2552.

[16] 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. DOI: 10.5664/jcsm.3630.

[17] H. Koch, J. A. Christensen, R. Frandsen, M. Zoetmulder, L. Arvastson, S. R. Christensen, P. Jennum, and H. B. Sorensen, “Automatic sleep classification using a data-driven topic model reveals latent sleep states,” *J. Neurosci. Methods*, vol. 235, pp. 130–137, 2014. DOI: 10.1016/j.jneumeth.2014.07.002.

[18] 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, 2017. DOI: 10.1109/TNSRE.2017.2721116.

[19] 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, 2018. DOI: 10.1109/TNSRE.2018.2813138.

[20] A. N. Olesen, P. Jennum, P. Peppard, E. Mignot, and H. B. D. Sorensen, “Deep residual networks for automatic sleep stage classification of raw polysomnographic waveforms,” in *2018 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Honolulu, HI, USA: IEEE, 2018, pp. 1–4. DOI: 10.1109/EMBC.2018.8513080.

[21] S. Biswal, H. Sun, B. Goparaju, M. B. Westover, J. Sun, and M. T. Bianchi, “Expert-level sleep scoring with deep neural networks,” *J. Am. Med. Informatics Assoc.*, vol. 25, no. 12, pp. 1643–1650, 2018. DOI: 10.1093/jamia/ocy131.

[22] J. B. Stephansen, A. N. Olesen, M. Olsen, A. Ambati, E. B. Leary, H. E. Moore, O. Carrillo, L. Lin, F. Han, H. Yan, Y. L. Sun, Y. Dauvilliers, S. Scholz, L. Barateau, B. Hogl, A. Stefani, S. C. Hong, T. W. Kim, F. Pizza, G. Plazzi, S. Vandi, E. Antelmi, D. Perrin, S. T. Kuna, P. K. Schweitzer, C. Kushida, P. E. Peppard, H. B. D. Sorensen, P. Jennum, and E. Mignot, “Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy,” *Nat. Commun.*, vol. 9, no. 1, p. 5229, 2018. DOI: 10.1038/s41467-018-07229-3.

[23] 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, 2019. DOI: 10.1109/TBME.2018.2872652.

[24] ——, “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, 2019. DOI: 10.1109/TNSRE.2019.2896659.

[25] A. N. Olesen, S. Chambon, V. Thorey, P. Jennum, E. Mignot, and H. B. D. Sorensen, “Towards a Flexible Deep Learning Method for Automatic Detection of Clinically Relevant Multi-Modal Events in the Polysomnogram,” in *2019 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Berlin, Germany: IEEE, 2019, pp. 556–561. DOI: 10.1109/EMBC.2019.8856570.

[26] D. Alvarez-Estevez and I. Fernández-Varela, “Large-scale validation of an automatic EEG arousal detection algorithm using different heterogeneous databases,” *Sleep Med.*, vol. 57, pp. 6–14, 2019. DOI: 10.1016/j.sleep.2019.01.025.

[27] A. Brink-Kjaer, A. N. Olesen, P. E. Peppard, K. L. Stone, P. Jennum, E. Mignot, and H. B. Sorensen, “Automatic Detection of Cortical Arousals in Sleep and their Contribution to Daytime Sleepiness,” *Clin. Neurophysiol.*, 2020. DOI: 10.1016/j.clinph.2020.02.027. arXiv: 1906.01700 [q-bio.NC].

[28] L. Carvelli, A. N. Olesen, A. Brink-Kjær, E. B. Leary, P. E. Peppard, E. Mignot, H. B. Sørensen, and P. Jennum, “Design of a deep learning model for automatic scoring of periodic and non-periodic leg movements during sleep validated against multiple human experts,” *Sleep Med.*, vol. 69, pp. 109–119, 2020. DOI: 10.1016/j.sleep.2019.12.032.

[29] J. B. Blank, P. M. Cawthon, M. L. Carrion-Petersen, L. Harper, J. P. Johnson, E. Mitson, and R. R. Delay, “Overview of recruitment for the osteoporotic fractures in men study (MrOS),” *Contemp. Clin. Trials*, vol. 26, no. 5, pp. 557–568, 2005. DOI: 10.1016/j.cct.2005.05.005.

[30] E. Orwoll, J. B. Blank, E. Barrett-Connor, J. Cauley, S. Cummings, K. Ensrud, C. Lewis, P. M. Cawthon, R. Marcus, L. M. Marshall, J. McGowan, K. Phipps, S. Sherman, M. L. Stefanick, and K. Stone, “Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study — A large observational study of the determinants of fracture in older men,” *Contemp. Clin. Trials*, vol. 26, no. 5, pp. 569–585, 2005. DOI: 10.1016/j.cct.2005.05.006.

[31] T. Blackwell, K. Yaffe, S. Ancoli-Israel, S. Redline, K. E. Ensrud, M. L. Stefanick, A. Laffan, and K. L. Stone, “Associations Between Sleep Architecture and Sleep-Disordered Breathing and Cognition in Older Community-Dwelling Men: The Osteoporotic Fractures in Men Sleep Study,” *J. Am. Geriatr. Soc.*, vol. 59, no. 12, pp. 2217–2225, 2011. DOI: 10.1111/j.1532-5415.2011.03731.x.

[32] A. Rechtschaffen and A. Kales, Eds., *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. Washington, DC: National Institute of Health, 1968.

[33] American Sleep Disorders Association, “EEG arousals: scoring rules and examples: a preliminary report from the Sleep Disorders Atlas Task Force of the American Sleep Disorders Association,” *Sleep*, vol. 15, no. 2, pp. 173–184, 1992, PMID: 11032543.

[34] M. Zucconi, R. Ferri, R. Allen, P. C. Baier, O. Bruni, S. Chokroverty, L. Ferini-Strambi, S. Fulda, D. Garcia-Borreguero, W. A. Hening, M. Hirshkowitz, B. Högl, M. Hornyak, M. King, P. Montagna, L. Parrino, G. Plazzi, and M. G. Terzano, “The official World Association of Sleep Medicine (WASM) standards for recording and scoring periodic leg movements in sleep (PLMS) and wakefulness (PLMW) developed in collaboration with a task force from the International Restless Legs Syndrome Study Grou,” *Sleep Med.*, vol. 7, no. 2, pp. 175–183, 2006. DOI: 10.1016/j.sleep.2006.01.001.

[35] D. A. Dean, A. L. Goldberger, R. Mueller, M. Kim, M. Rueschman, D. Mobley, S. S. Sahoo, C. P. Jayapandian, L. Cui, M. G. Morrical, S. Surovec, G.-Q. Zhang, and S. Redline, "Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource," *Sleep*, vol. 39, no. 5, pp. 1151–1164, 2016. DOI: 10.5665/sleep.5774.

[36] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline, "The National Sleep Research Resource: towards a sleep data commons," *J. Am. Med. Informatics Assoc.*, vol. 25, no. 10, pp. 1351–1358, 2018. DOI: 10.1093/jamia/ocy064.

[37] Y. A. LeCun, L. Bottou, G. B. Orr, and K. R. Müller, "Efficient backprop," in *Neural Networks: Tricks of the Trade*, ser. Lect. Notes Comput. Sci. vol 7700, G. Montavon, G. B. Orr, and K.-R. Müller, Eds., Springer Berlin Heidelberg, 2012. DOI: 10.1007/978-3-642-35289-8-3.

[38] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in *Proc. 27th Int. Conf. Mach. Learn.*, Haifa, Israel, 2010.

[39] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, Lille, France: JMLR, 2015. arXiv: 1502.03167 [cs.LG].

[40] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches," in *Proc. SSST-8, Eighth Work. Syntax. Semant. Struct. Stat. Transl.*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 103–111. DOI: 10.3115/v1/W14-4012.

[41] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," 2014. arXiv: 1412.3555 [cs.NE].

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *31st Conf. Neural Inf. Process. Syst. (NIPS 2017)*, Long Beach, CA, USA, 2017. arXiv: 1706.03762 [cs.CL].

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019. arXiv: 1810.04805v2 [cs.CL].

[44] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *3rd Int. Conf. Learn. Represent. (ICLR 2015)*, San Diego, CA, USA, 2015. arXiv: 1409.0473 [cs.CL].

[45] A. N. Olesen, P. Jennum, E. Mignot, and H. B. D. Sorensen, "Deep transfer learning for improving single-EEG arousal detection," 2020.

[46] S. Chambon, V. Thorey, P. J. Arnal, E. Mignot, and A. Gramfort, "A Deep Learning Architecture to Detect Events in EEG Signals During Sleep," in *2018 IEEE 28th Int. Work. Mach. Learn. Signal Process.*, Aalborg, Denmark: IEEE, 2018, pp. 1–6. DOI: 10.1109/MLSP.2018.8517067.

[47] S. Chambon, V. Thorey, P. Arnal, E. Mignot, and A. Gramfort, "DOSED: A deep learning approach to detect multiple sleep micro-events in EEG signal," *J. Neurosci. Methods*, vol. 321, pp. 64–78, 2019. DOI: 10.1016/j.jneumeth.2019.03.017.

[48] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," in *Comput. Vis. – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham, Switzerland: Springer, 2016, pp. 21–37. DOI: 10.1007/978-3-319-46448-0_2.

[49] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, 2020. DOI: 10.1109/TPAMI.2018.2858826.

[50] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2014. arXiv: 1412.6980 [cs.LG].

G

POPULAR SCIENCE ARTICLE

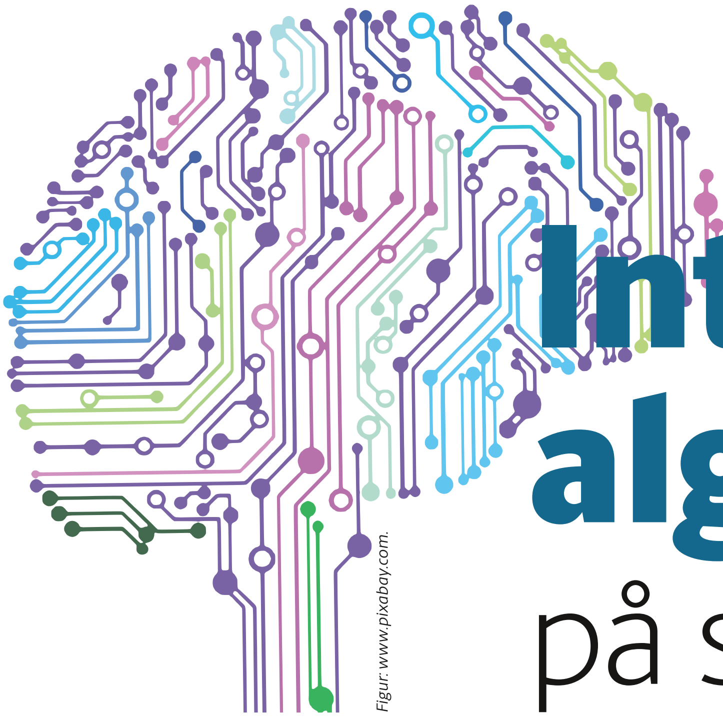
TITLE: Intelligente algoritmer på søvnklinikken

AUTHORS: Alexander Neergaard Olesen

PUBLICATION: Medicoteknik - Magasin for Dansk Medicoteknisk Selskab

STATUS: Published

FULL CITATION: A. N. Olesen, "Intelligente algoritmer på søvnklinikken," Medicoteknik, no. 2, pp. 25–27, April, 2020 [Online]. Available: http://ipaper.ipapercms.dk/TechMedia/Medicoteknik/2020/?page=24. [Accessed: April 15, 2020].



Figur: www.pixabay.com.

Intelligente algoritmer på søvnklinikken

Nyudviklede algoritmer kan assistere sundhedsfagligt personale på søvnklinikker. Gevinsten er hurtigere og mere konsistente diagnoser - og dermed bedre behandling af søvnsygdomme og andre lidelser med udtalt søvnbesvær.

Af Alexander Neergaard Olesen. Ph.d.-studerende – Institut for Sundhedsteknologi, DTU

Selv om vi i snit bruger en tredjedel af vores liv på at sove, er forskning i søvn og søvnmedicinske sygdomme stadig et relativt nyt felt. Først i slutningen af 60'erne blev de første komplette retningslinjer for klinisk analyse af søvnmønstre etableret.

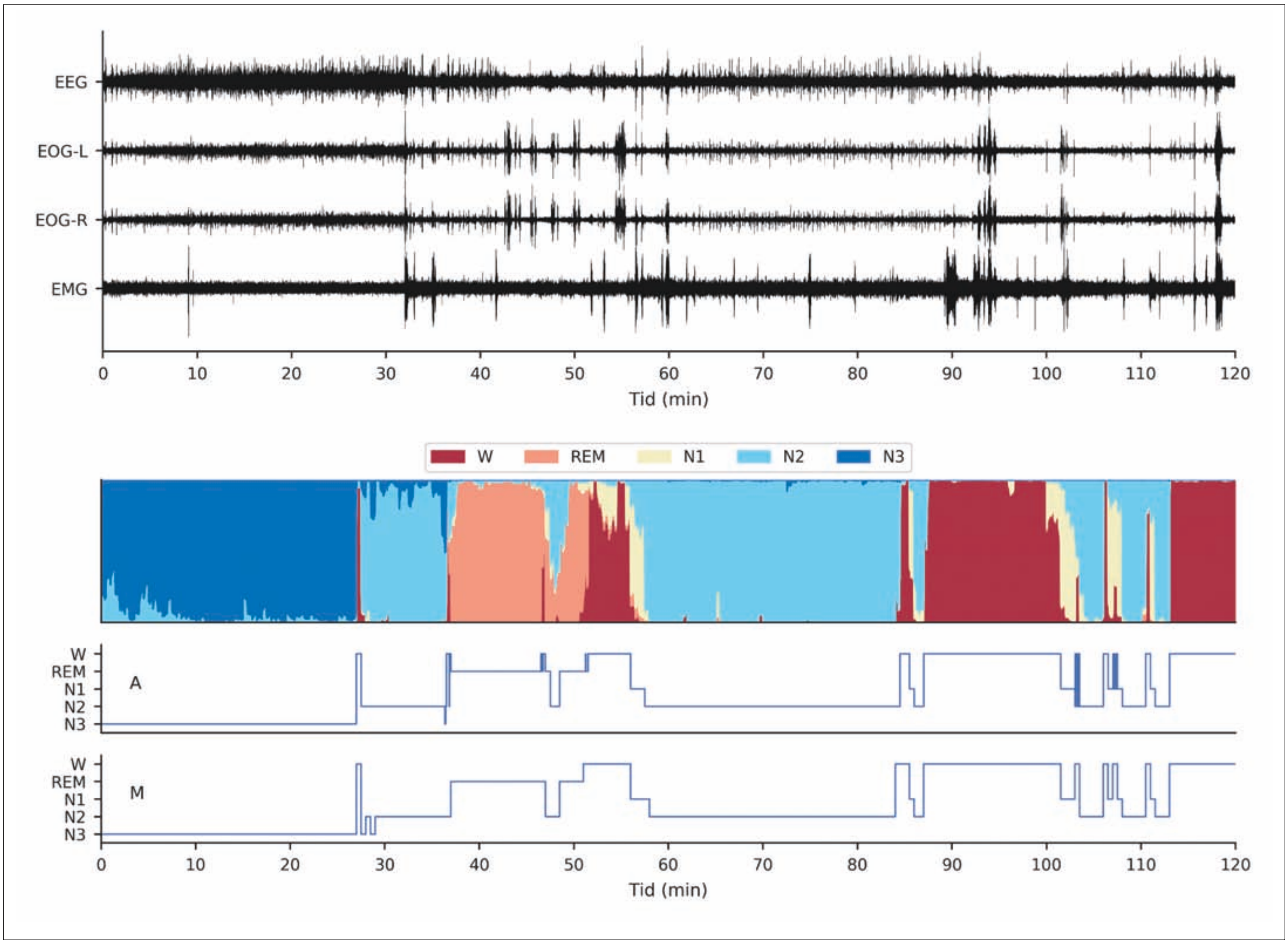
Og trods stigende interesse i søvn fra et samfunds- og sundhedsvidenskabeligt synspunkt er der stadig mange ubesvarede spørgsmål om de underliggende mekanismer i hjernen: Hvordan vi opnår gode søvnvaner, og hvordan vi skal forholde os til og behandle søvnlidelser.

I min forskning har jeg fokuseret på at udvikle intelligente supportsystemer, som kan bruges af teknikere og søvnlæger i en klinisk sammenhæng.

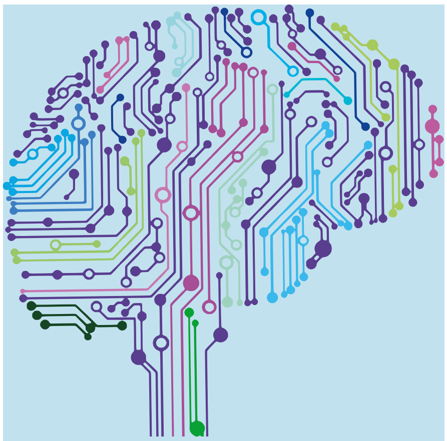
Maskinerne træder ind i klinikken

Når patienter henvises til en søvnklinik, vil de typisk blive undersøgt med en såkaldt polysomnografi (PSG). Det er en samlet betegnelse for optagelse af hjerne-, hjerte-, øjen-, respirations- og muskelfunktion under søvn. Disse opta-

▶



Figur 1. Her ses et eksempel på 120 minutters polysomnografi-data (elektro-encefalografi, EEG, venstre/højre elektro-okulografi, EOG, elektromyografi, EMG), der gennem vores søvnscoringsalgoritme omdannes til dels en hypnodensitet (farvet) og et automatisk (A) og manuelt (M) registreret hypnogram.



Om projektet

Forskningsprojektet er en del af et større nordatlantisk samarbejde mellem DTU Sundhedsteknologi, Rigshospitalet og Stanford University. Kontaktpersoner: lektor Helge B. D. Sørensen, DTU Sundhedsteknologi; professor Poul Jørgen Jennum, Rigshospitalet; professor Emmanuel Mignot, Stanford University.

gelser bliver derefter manuelt undersøgt og analyseret af specialister i søvnanalyse ud fra specifikke retningslinjer, som er udarbejdet af American Academy of Sleep Medicine.

Afhængigt af situationen skal der registreres søvnstadier, og der skal muligvis også annoteres områder med korte opvågninger, benspjæt, apnø-anfald og desaturationer (korte perioder med for lav iltmætning i blodet). Dette kan tage flere timer for en specialist at udføre. Derudover har flere studier vist, at specialisterne ikke altid er enige i analyserne. Selv den samme specialist vil ikke registrere den samme PSG på præcis samme måde hver gang.

Derfor er der et stort behov for robuste metoder til at assistere dette arbejde for at sikre en præcis og konsekvent analyse af søvnen hver gang.

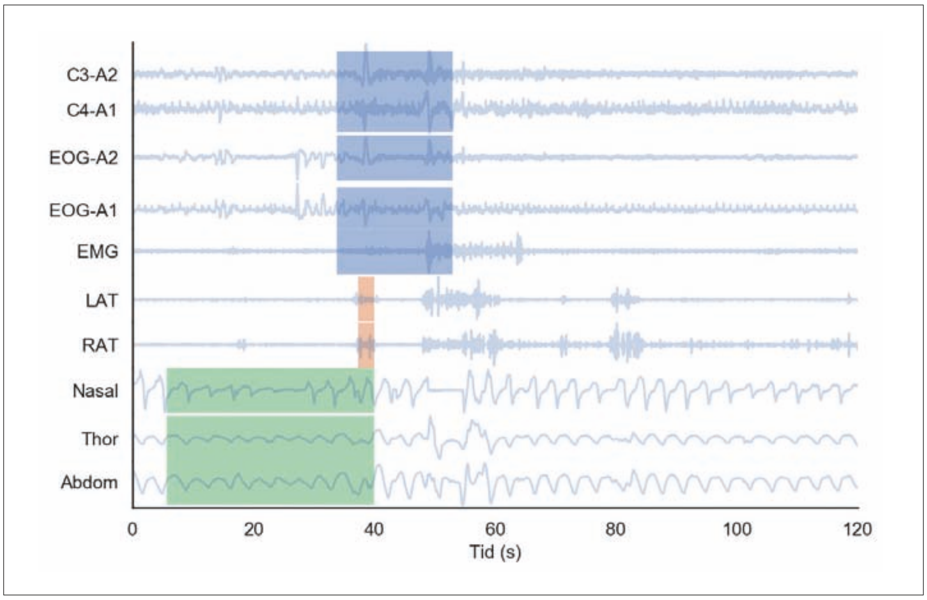
Rå signaler

Et gennemgående element i min forskning har været at modellere og analysere de rå PSG-signaler i stedet for andre repræsentationer af data som frekvensspektra og spektrogrammer, som hidtil ofte er blevet anvendt. Hypotesen er, at denne form for behandling kan introducere et uønsket menneskeligt bias i repræsentationen af data, som kan skjule vigtige underliggende mønstre. Eksempelvis har jeg i samarbejde med mine vejledere udviklet en model til at klassificere søvnstadier baseret på de rå PSG-signaler fra hjernen, øjne og muskler under hagen. Progressionen af søvnstadierne sammenfattes i et hypnogram, der blandt andet illustrerer, hvordan søvnstadierne gentages i cyklusser over natten.

Vi har i vores gruppe også forsket i at beskrive hypnogrammet ved hjælp af sandsynlighedsfunktioner (hypnodensitet), som kan give et mere detaljeret indblik i dynamikken i hjernen.

Dette har vi brugt til at udvikle et fuldautomatisk diagnostisk værktøj, som ud fra blot en enkelt nats optagelser kan hjælpe med at bestemme, om patienten

Figur 2. Et eksempel på, hvad mikro-event-modellen kan detektere. Her ses en periode på ca. 35 sekunder med vejrtrækningsbesvær (grønt markeret) i de kanaler, der måler tryk i næsen (Nasal), brystekspansion (Thor) og maveekspansion (Abdo). Herefter ses en »arousal« i hjerne- (EEG-A1/A2), øjen- (EOG-A1/A2) og muskelsignalerne (EMG) fra omkring 35 s (blåt markeret) samt et kort benspjæt (orange markeret) lige inden 40 s markøren på venstre og højre ben (LAT/RAT). Denne sekvens med vejrtrækningsbesvær samt arousal og/eller benspjæt ses ofte sammen.



lider af narkolepsi. Typisk vil denne gruppe af patienter skulle undergå flere PSG-optagelser samt en opfølgende multipel søvnlatenstest (MSLT) og/eller lumbalpunktur mm., før en endelig diagnose kan stilles.

Mod automatisk analyse

Udover at bestemme og beskrive søvnarkitekturen ved registrering af søvnstadier findes der også andre aspekter af søvn, som har klinisk relevans.
Jeg har i min forskning beskæftiget mig med detektion og annotation af »mikro-events«, såsom »arousals« (korte opvågninger under søvn), regelmæssige benspjæt og perioder med udtalt vejrtrækningsbesvær. Sammen med et hold forskere fra Frankrig har jeg udviklet en model baseret på deep learning, der, ligesom søvnstadiemodellen, kan analysere et sæt rå data fra en PSG og automatisk finde de områder, hvor disse events opstår.
Det smarte ved denne model er, at den er meget fleksibel i forhold til, hvad man specifikt er interesseret i. Modellen er da også blevet brugt i forbindelse med detektion af søvnspindler og K-komplekser, som er nogle meget specifikke hjernebølger, der typisk ses i bestemte søvnstadier.
Udviklingen af denne model er et skridt på vejen mod et automatisk, diagnostisk supportsystem til generel analyse af søvnstudier, som vi arbejder på i vores forskningsgruppe.

Datamængden er afgørende

Disse specifikke modeller baseret på deep learning stiller store krav til mængden og variabiliteten af tilgængelige data, da de som oftest består af mange millioner parametre. En enkelt af vores modeller består eksempelvis af ca. 30 millioner parametre, hvilket dog er relativt beskedent i forhold til, hvad de førende industrielle forskningsgrupper i Google og Facebook arbejder med.
I vores forskningsgruppe har vi derfor gennem et frugtbart internationalt samarbejde med Stanford Center for Sleep Sciences and Medicine og Dansk Center for Søvnmedicin indsamlet flere tusinde søvnstudier til vores forskning i intelligente medicinske support-systemer.
Det har blandt andet ført til et studie, hvor vi har undersøgt, hvordan forskellige datasæt påvirker modellernes evne til at generalisere til nye data. Dette er vigtigt at undersøge, fordi en model trænet på ét datasæt med en specifik patientgruppe højst sandsynligt ikke virker efter hensigten på et andet datasæt med helt andre patienter. Det kan skyldes, at en model bliver trænet på raske subjekter med normale søvnmønstre - og derefter benyttes på patienter med Parkinsons sygdom eller en anden neurodegenerativ sygdom, der påvirker kontrolcentrene i hjernestammen, som styrer søvnen. Hvis en algoritme eller model ikke bliver vist eksempler på disse søvnmønstre under træningen, kan den ikke genkende dem ordentligt.

Kan vi stole på algoritmerne?

I vores forskningsgruppe har vi skarpt fokus på udvikling af robuste algoritmer, der kan benyttes i kliniske sammenhænge af lægefagligt personale. For eksempel har vi i flere af vores studier testet vores algoritmer op mod en konsensus af adskillige søvnspecialister for at sikre, at vores modeller virker og er konsistente.
Systemer baseret på deep learning-algoritmer bliver ofte mødt med skepsis. Kritikere påpeger, at algoritmerne er såkaldte »black boxes«, hvor vi i virkeligheden ikke ved, hvad der ligger til grund for en specifik beslutning. Flere forskergrupper har dog undersøgt metoder til at »åbne op for kassen« - et felt, der populært betegnes som »explainable AI«.
Min - og mine vejlederes - vision er, at søvnklinikker i fremtiden kan bruge vores systemer til dels at fremme ny forskning i søvn, men vigtigst af alt: At patienter med søvnlidelser kan få bedre, hurtigere og mere præcis afklaring af deres søvnproblemer, hvilket i sidste ende vil føre til en bedre behandling.
Det er dog vigtigt at understrege, at vi på ingen måde forestiller os, at de intelligente systemer skal erstatte hverken lægefagligt eller teknisk personale. Tværtimod ser vi vores forskning som værktøjer, der kan assistere personalet og effektivisere deres arbejde og hverdag - til gavn for patienterne.