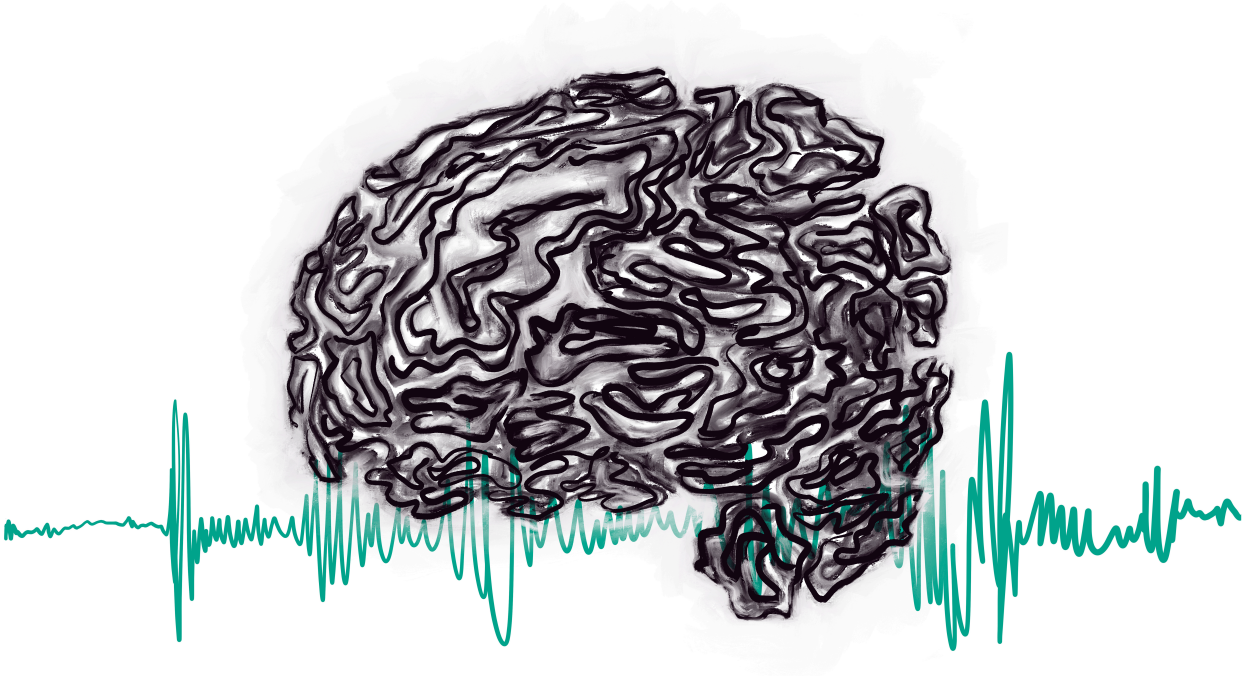
UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE



PhD thesis

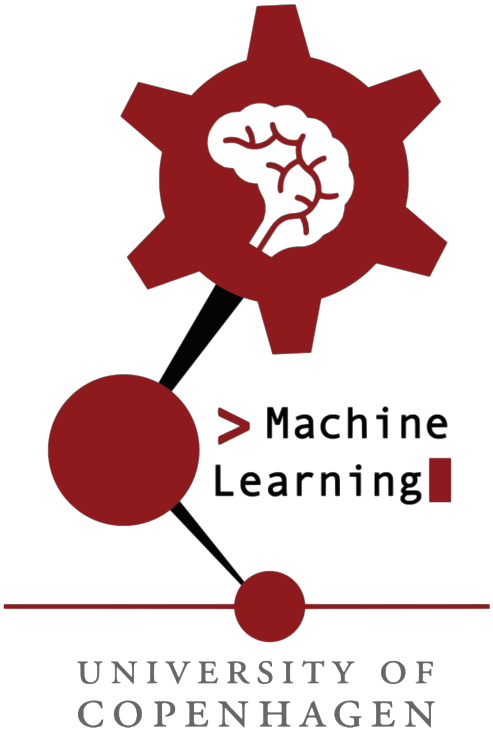
Segmentation of medical images and time series using fully convolutional neural networks

Mathias Perslev

Supervisor: Christian Igel

This thesis has been submitted to the PhD School of the Faculty of Science, University of Copenhagen
May 9, 2023

Submitted	May 9, 2023
Institution	University of Copenhagen
Department	Department of Computer Science
Section	Machine Learning Section
Supervisor	Prof. Christian Igel Department of Computer Science University of Copenhagen
Clinical co-supervisor	MD, Prof. Poul Jørgen Jennum Danish Center for Sleep Medicine Rigshospitalet, RegionH
Assessment committee	*Chair* Assoc. Prof. Melanie Ganz-Benjaminsen Department of Computer Science University of Copenhagen *Opponents* Prof. Klaus-Robert Müller Maschinelles Lernen Technische Universität Berlin Prof. M. Brandon Westover Beth Israel Deaconess Medical Center Harvard Medical School
Funding	Independent Research Fund Denmark (DFF) Grant *U-Sleep*. Project number 9131-00099B.
Thesis cover art	Jeanne Vuaille



Preface

The work presented in this thesis was conducted in the Machine Learning Section of the Department of Computer Science, University of Copenhagen, between September 2019 and May 2023. Part III of the thesis was done in collaboration with the Danish Center for Sleep Medicine at Rigshospitalet, Glostrup, Denmark.

The thesis extends on preliminary work conducted at the Department of Computer Science, University of Copenhagen, between January 2018 and August 2019, as part of my master's thesis and subsequent Research Assistant position. For completeness, manuscripts based on this work were included in the thesis if published between September 2019 and May 2023.

The thesis was supervised by Professor Christian Igel from the Department of Computer Science, University of Copenhagen (Parts II and III). Part III was co-supervised by MD Professor Poul Jørgen Jennum from Rigshospitalet, RegionH.

In the following, I will use *we*, *our*, etc., to collectively refer to myself, my supervisors and all other colleagues who contributed to the findings of this thesis.

Acknowledgements

This thesis is the culmination of an exciting research journey that began with my Master's thesis several years ago. Along the way, I have received invaluable support from several colleagues, family members, and friends.

First and foremost, I would like to thank my supervisor, **Christian Igel**, for giving me the autonomy to work independently and explore scientific ideas of my interest, yet always being available with a minute's notice for feedback, guidance, and emotional support. I am grateful for our collaboration and look forward to continuing it in the years to come.

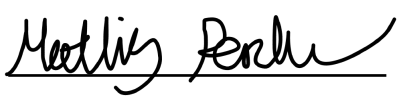
I also thank my clinical co-supervisor, **Poul Jennum**, for sharing your extensive knowledge and enthusiasm for sleep medicine and its technological advancements. I am eager to continue working together to implement our developed sleep staging models clinically. Thank you to all other collaborators from the Danish Center for Sleep Medicine for sharing data and offering valuable insights. Thank you **Miki Nikolic** and **Lykke Kempfner** for engaging in discussions and providing feedback on manuscripts and abstracts.

I am grateful to my excellent assessment committee for dedicating their time and effort to evaluate my thesis, and I look forward to discussing it.

I would also like to thank **Akshay Pai**, **Sune Darkner**, and **Erik B. Dam**, who introduced me to the exciting world of machine learning and radiology during my Master's thesis and early PhD. Cheers to **Andreas Lauritzen**, **Rasmus Kær Jørgensen**, and **Svetlana Kutuzova** for the many enjoyable conversations, social gatherings, and coffee breaks. Thank you to all my other great colleagues at DIKU. Thank you, **Anne Marie Mira Lindegaard**, for coffee in a critical moment.

I am incredibly fortunate to have had the ever support of my family, friends (including those who questioned whether I would ever complete my thesis), and the fantastic **Jeanne Vuaille** throughout my PhD journey. This thesis and the time making it would not have been the same without you.

It is with great excitement that I present this thesis, and I hope you find it interesting.



Mathias Perslev

Copenhagen, May 8, 2023

Enclosed publications and manuscripts

Mathias Perslev, Erik B. Dam, Akshay Pai, and Christian Igel. One Network To Segment Them All: A General, Lightweight System for Accurate 3D Medical Image Segmentation. Medical Image Computing and Computer Assisted Intervention (MICCAI), LNCS 11765, pp. 30-38, Springer, 2019.
https://doi.org/10.1007/978-3-030-32245-8_4

Mathias Perslev, Akshay Pai, Jon Runhaar, Christian Igel, and Erik B. Dam. Cross-Cohort Automatic Knee MRI Segmentation with Multi-Planar U-Nets. Journal of Magnetic Resonance Imaging 55(6):1650-1663, 2022.
https://doi.org/10.1002/jmri.27978

Mathias Perslev, Michael Hejselbak Jensen, Sune Darkner, Poul Jørgen Jennum, and Christian Igel. U-Time: A Fully Convolutional Network for Time Series Segmentation Applied to Sleep Staging. Advances in Neural Information Processing Systems (NeurIPS 2019), pp. 4417-4428, 2019.
https://arxiv.org/abs/1910.11162

Mathias Perslev, Sune Darkner, Lykke Kempfner, Miki Nikolic, Poul Jørgen Jennum, and Christian Igel. U-Sleep: Resilient High-Frequency Sleep Staging. npj Digital Medicine 4, 2021.
https://doi.org/10.1038/s41746-021-00440-5

Mathias Perslev, Anders Sode West, Sofie Amalie Simonsen, Laura Bødker Ponsaing, Helle Klingenberg Iversen, Christian Igel, and Poul Jørgen Jennum. Automatic detection of abnormal sleeping patterns in stroke patients using high-frequency sleep staging. Journal of Sleep Research 31(S1), Oral Abstract, 2022.
https://doi.org/10.1111/jsr.13739

Mathias Perslev, Shaun Purcell, Miki Nikolic, Lykke Kempfner, Poul Jørgen Jennum, and Christian Igel. U-Sleep v2: Single-Channel, High-Frequency, and Spatial Sleep Staging for Complex EEG.
In preparation

Related work not included in the thesis

Michela Antonelli, Annika Reinke, Spyridon Bakas et al. The Medical Segmentation Decathlon. Nature Communications 13, 4128 (2022). https://doi.org/10.1038/s41467-022-30695-9

Patrick Bilic, Patrick Christ, Hongwei Bran Li et al. The Liver Tumor Segmentation Benchmark (LiTS). Medical Image Analysis 84, 2023. https://doi.org/10.1016/j.media.2022.102680

Arjun D. Desai, Francesco Caliva, Claudia Iriondo et al. The International Workshop on Osteoarthritis Imaging Knee MRI Segmentation Challenge: A Multi-Institute Evaluation and Analysis Framework on a Standardized Dataset. Radiology: Artificial Intelligence 3:e200078, 2021. https://doi.org/10.1148/ryai.2021200078

Erik B. Dam, Arjun D. Desai, Cem M. Deniz et al. Towards Automatic Cartilage Quantification in Clinical Trials – Continuing from the 2019 IWOAI Knee Segmentation Challenge. Osteoarthritis Imaging, 100087, 2023. https://doi.org/10.1016/j.ostima.2023.100087

Thorbjørn Louring Koch, Mathias Perslev, Christian Igel, and Sami Sebastian Brandt. Accurate Segmentation of Dental Panoramic Radiographs with U-Nets. IEEE International Symposium on Biomedical Imaging (ISBI), pp. 15-19, IEEE Press, 2019. https://doi.org/10.1109/ISBI.2019.8759563

Sandeep Singh Sengar, Christopher Meulengracht, Mikael Ploug Boesen, Anders Føhrby Overgaard, Henrik Gudbergsen, Janus Damm Nybing, Mathias Perslev, and Erik Bjørnager Dam. Multi-planar 3D knee MRI segmentation via UNet inspired architectures. International Journal of Imaging Systems and Technology, pp. 1- 14, 2022. https://doi.org/10.1002/ima.22836

William Michael Laprade, Mathias Perslev & Jon Sporring. How Few Annotations are Needed for Segmentation Using a Multi-planar U-Net? Deep Generative Models, and Data Augmentation, Labelling, and Imperfections. DGM4MICCAI DALI. Lecture Notes in Computer Science, vol 13003. Springer, Cham, 2021. https://doi.org/10.1007/978-3-030-88210-5_20

Contents

Abstract

Diagnostic tasks in healthcare often involve segmenting regions of interest in images and time series, such as outlining organs in medical scans or scoring physiological events in electroencephalography (EEG) recordings. Medical professionals perform most of these complex and time-consuming tasks manually, leading to potential errors and limiting diagnostic efficiency. With the increasing global diagnostic burden on healthcare systems, there is a growing need for (semi-) automatic computer systems to alleviate repetitive manual tasks. Furthermore, these systems can make expert knowledge available for people with limited access to well-trained medical doctors.

The primary aim of this thesis was to develop clinically robust automatic segmentation systems for medical images and time series based on recent advances in machine learning. The thesis comprises two parts.

The first part focused on developing a machine learning model for general medical 3D image segmentation, applicable across scanning modalities and tasks. We introduced the Multi-Planar U-Net, a fully convolutional neural network based on the U-Net architecture, which uses a data-augmentation scheme to resample randomly rotated 2D input images from 3D training data. This process enforces rotational equivariance properties and enables segmenting new scans from multiple orientations for ensemble-like predictions. The Multi-Planar U-Net demonstrated applicability to variable tasks in magnetic resonance (MR) and computerized tomography (CT) images without manual hyperparameter adjustments and proved competitive in multiple segmentation challenges, including the 2018 Medical Segmentation Decathlon and the 2020 International Workshop on Osteoarthritis Imaging Knee MRI Segmentation Challenge.

The second part of this thesis addressed the problem of automatic sleep staging of polysomnographic data, which involves segmenting physiological signals from sleeping individuals into distinct sleep stages. We developed a U-Net-based model for medical time series data called U-Time that leverages the similarities between sleep staging and image segmentation. This model outperformed alternative models typically used for automatic sleep staging and was transferable across clinical cohorts without hyperparameter re-tuning.

This ability inspired the development of a successor model, U-Sleep, designed for robust sleep staging on diverse polysomnography data. Trained on over 15,000 participants from 12 clinical studies, U-Sleep demonstrated expert-level accuracy and adaptability to different EEG input derivations, patient demographics, and recording equipment. It was also accurate for patients with severe brain disorders, such as stroke and Parkinson's disease, despite their absence in the training dataset. Fur-

thermore, we explored U-Sleep's ability to score sleep stages at higher-than-usual frequencies, which facilitated the separation of patients with sleep disorders or acute stroke from control groups, indicating potential biomarker development. Sleep metrics derived from U-Sleep's high-frequency sleep scores were more consistent than those from low-frequency and human expert scores, suggesting improved diagnostic accuracy. Consequently, U-Sleep may be a candidate for clinical sleep staging and a potential research tool for high-frequency sleep patterns. The model is available for research at `https://sleep.ai.ku.dk/`, where it has scored over 45,000 sleep studies.

In summary, this thesis presented clinically robust and accurate machine learning models for segmenting medical image volumes and time series. Key practices for developing such models were identified: First, we reconfirmed that fully convolutional, feed-forward-only neural networks like the U-Net are broadly applicable as they performed well across diverse tasks in medical images and time series. Second, we found it beneficial to design data-augmentation techniques that induce various model invariance or equivariance properties to input data transformations that increase clinical robustness, even if the target function becomes more complex, as long as the augmentations also significantly expand the set of actual training examples. Finally, we found clinical robustness achievable by training machine learning models on extensive and highly variable training datasets from multiple sources, even if datasets differ in recording hardware, patient population, or data preprocessing pipeline.

Dansk resumé

En række diagnostiske opgaver inden for sundhedsvæsenet indebærer segmentering af interesseområder i billeder og tidsserier. Eksempler inkluderer afgrænsning af organer eller tumorer i medicinske scanninger samt scoring af diverse fysiologiske hændelser i elektroencefalografi-optagelser (EEG). Medicinske fagfolk udfører de fleste af disse komplekse og tidskrævende segmenteringsopgaver manuelt, hvilket begrænser den diagnostiske proces. Da sundhedssystemerne oplever en stigende diagnostisk byrde, vil der i fremtiden være et øget behov for (semi-)automatiske computersystemer til at afhjælpe repetitive manuelle segmenteringsopgaver. Sådanne systemer kan desuden gøre ekspertviden tilgængelig for folk med begrænset adgang til veluddannede læger.

Det primære formål med denne afhandling var at udvikle klinisk robuste automatiske segmenteringssystemer til medicinske billeder og tidsserier baseret på nylige fremskridt inden for maskinlæringsteknologi. Afhandlingen består af to dele.

Den første del af afhandlingen fokuserer på udviklingen af en maskinlæringsmodel til segmentering af generelle medicinske 3D-billeder, som kan anvendes på tværs af segmenteringsopgaver og scanningsudstyr. Vi introducerer en model, kaldet *Multi-Planar U-Net*, som er et såkaldt *fully convolutional neural network* baseret på U-Net-arkitekturen, og som anvender en data-augmenteringsmekanisme til at udtrække tilfældigt roterede 2D-billeder fra et 3D datasæt, som bruges til at træne modellen. Denne proces introducerer rotationsækvivalens og muliggør segmentering af nye scanninger fra flere orienteringer, som efterfølgende kan kombineres til en enkelt og mere præcis segmentering. Multi-Planar U-Net har en tilpasningsevne, der gør modellen i stand til at segmentere både MR- og CT-scanninger (magnetisk resonans og computertomografi) uden manuelle hyperparameterjusteringer. Modellen blev fundet konkurrencedygtig i flere segmenteringsudfordringer, herunder *Medical Segmentation Decathlon* i 2018 og *International Workshop on Osteoarthritis Imaging Knee MRI Segmentation Challenge* i 2020.

Den anden del af afhandlingen omhandler automatisk klassificering af søvnstadier i polysomnografiske data, hvilket indebærer en segmentering af søvnstadier i fysiologiske signaler fra sovende individer. Vi introducerer endnu en U-Net-baseret model, kaldet U-Time, som er specialiseret til medicinske tidsserier, og som udnytter lighederne mellem søvnstadieklassificering og billedsegmentering. Denne model klarede sig bedre end alternative modeller, der typisk anvendtes til denne opgave, og modellen kunne overføres til nye kliniske kohorter uden manuel justering af modellens hyperparametre.

Denne evne inspirerede udviklingen af en opfølgende model, kaldet U-Sleep, der er designet til

robust klassificering af søvnstadier i klinisk polysomnografidata af høj variabilitet. U-Sleep blev trænet på over 15.000 forsøgspersoner fra 12 kliniske studier og udviste en nøjagtighed på niveau med menneskelige eksperter og en evne til at tilpasse sig input data fra meget forskellige EEG-kilder og optageudstyr og fra patienter af variabel demografi. Modellen blev også fundet nøjagtig på data fra patienter med alvorlige hjernesygdomme, såsom apopleksi og Parkinsons sygdom, på trods af fraværet af data fra sådanne patienter i træningsdatasættet. Vi udforskede desuden U-Sleeps evne til at score søvnstadier ved højere end sædvanlige frekvens. Disse højfrekvente søvnstadier gjorde det lettere at adskille patienter med søvnforstyrrelser eller apopleksi fra kontrolgrupper, hvilket indikerer et potentiale for biomarkørudvikling baseret på U-Sleep modellen. Søvnmetrikker afledt af U-Sleeps højfrekvente søvnscorer var desuden mere konsistente end dem afledt fra lavfrekvente scorer, hvilket indikerer at U-Sleep kan forbedre den diagnostiske nøjagtighed. U-Sleep er samlet set en lovende kandidat til klinisk klassificering af søvnstadier og et potentielt forskningsværktøj, som kan give indsigt i højfrekvente søvnmønstre. Modellen er frit tilgængelig for forskning på `https://sleep.ai.ku.dk/`, hvor den har scoret mere end 45.000 søvnundersøgelser til dato.

Sammenfattende præsenterer denne afhandling klinisk robuste og præcise maskinlæringsmodeller til segmentering af medicinske 3D-billeder og tidsserier. En række metoder til udvikling af sådanne modeller blev identificeret: For det første blev det bekræftet, at *fully convolutional*, *feed-forward-only* neurale netværk, som f.eks. U-Net, er bredt anvendelige, da de klarede sig godt på tværs af forskellige opgaver i både medicinske billeder og tidsserier. For det andet fandt vi det fordelagtigt at designe data-augmenteringsmekanismer, der inducerer invarians- eller ækvivalens over for diverse transformationer af input data for at øge modellens kliniske robusthed, selv hvis den opgave, modellen skal løse, bliver mere kompleks, så længe augmenteringerne også udvider sættet af faktiske træningseksempler betydeligt. Endelig fandt vi, at klinisk robusthed kan opnås ved at træne maskinlæringsmodeller på omfattende og meget variable træningsdatasæt fra flere kilder, selv om datasættene adskiller sig fra hinanden med hensyn til optagelsesudstyr, patientdemografi eller præprocessering af data.

Overview of thesis structure

This thesis concerns the development of clinically robust machine learning models for segmenting medical images and time series. It is split into four parts: Part I introduces the motivation, main objectives, scientific background, and summaries of all enclosed papers and manuscripts. Part II contains work on medical image segmentation using fully convolutional neural networks. Part III considers time series segmentation using similar models applied to sleep staging. Part IV discusses the findings of Parts II and III, highlights study limitations, draws overall conclusions, and provides an outlook for future work and best practices for developing robust machine learning models for healthcare applications.

Part I

Introduction

Chapter 1

Motivation

1.1 The burden of manual segmentation

Many diagnostic tasks in healthcare require segmenting regions of interest in images and time series (Faust, Hagiwara, et al. 2018; Simpson et al. 2019). Segmentation is the process of dividing an input into distinct areas. Examples include outlining organs and lesions in medical scans and scoring physiological events like sleep stages or seizures in electroencephalography (EEG) recordings. See Figure 1.1 for visual examples of medical image and time series segmentation problems. Segmentation allows quantification of, for instance, a brain tumour's existence, volume and location, from which a diagnosis or radiation plan may be derived (Assefa et al. 2010; ICRU 1999; Menze et al. 2014). Medical doctors or technicians perform most segmentation tasks through complicated and time-consuming manual inspections. These are often expensive and error-prone with significant inter-rater variability, thus limiting diagnostic throughput and precision (Danker-Hopfe, Anderer, et al. 2009; Joskowicz et al. 2019). A few examples are:

- Segmentation of *gliomas* (a common primary brain tumour) in computed tomography (CT) or magnetic resonance imaging (MRI) is becoming common to track the development of tumour size and morphology (Bauer et al. 2013). The segmentation is difficult due to the high variability in tumour size, location and appearance, often low contrast between the tumour and surrounding tissue and because the growing tumour may alter sounding structures making it more difficult to rely on knowledge of typical brain anatomy (Menze et al. 2014). Consequently, manual glioma segmentation is time-consuming and has high inter-rater variability (Angelini et al. 2007; Deeley et al. 2011; Weltens et al. 2001).

- Accurate segmentation of *liver tumours* in CT is a prerequisite for both diagnosis and treatment because the diameter of the lesion must be measured under the modified Response Evaluation Criteria in Solid Tumor (RECIST, Eisenhauer et al. 2009) guidelines to assess tumour burden and the exact tumour location known for effective treatment with, for instance, thermal ablation or radiotherapy (Albain et al. 2009; Shiina et al. 2018). However, the segmentation of liver tumours is complex because of the often low and variable contrast between the surrounding liver tissue and lesions, significant variability in tumour size, shape and location, and the possibility of multiple types of tumours co-occurring. Hence, manual segmentation is time-consuming and

subject to inter-rater variability (Bilic et al. 2023; Moghbel et al. 2018).

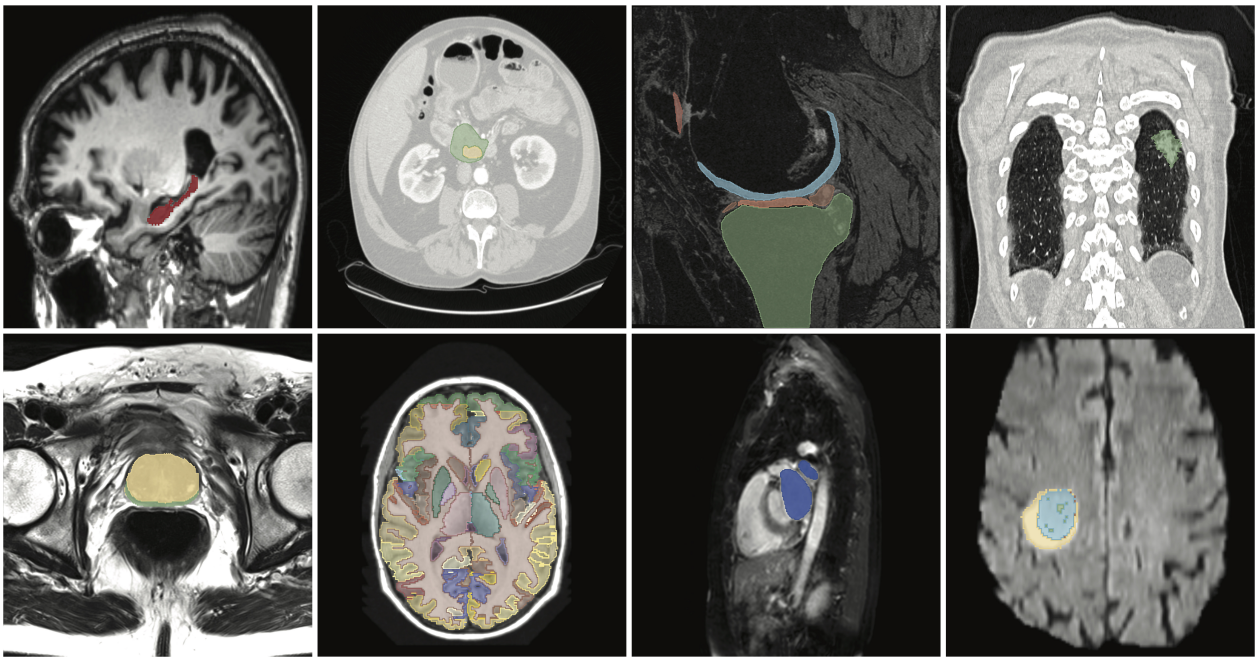
- Segmentation of *ventricles, atria and great vessels* in cardiovascular MRI forms the basis for diagnosing many functional and structural cardiovascular diseases (Myerson et al. 2010). This task is also considered highly time-consuming, and the manual segmentation suffers from high intra- and inter-rater variability (Peng et al. 2016). In addition, the number of cardiac MRIs conducted is expected to increase due to the growing prevalence of cardiovascular disease, which is projected to cause more than 23 million deaths globally by 2030 (Mathers et al. 2006).

- Similarly, segmentations are often needed for (physiological) time series recordings. A standard task in sleep medicine is *sleep staging* in which distinct physiological stages of sleep are segmented in polysomnography (PSG) data (a sleep study recording modalities such as EEG). Sleep staging forms the basis of many diagnostic tasks in sleep medicine but takes multiple hours per patient and suffers from high inter-rater variability (Danker-Hopfe, Anderer, et al. 2009; Rosenberg et al. 2013; Younes, Kuna, et al. 2018; Younes, Raneri, et al. 2016; X. Zhang et al. 2015). Automating and improving sleep staging is the focus of Part III of this thesis. For an extended introduction to sleep staging, see Chapter 3.3.

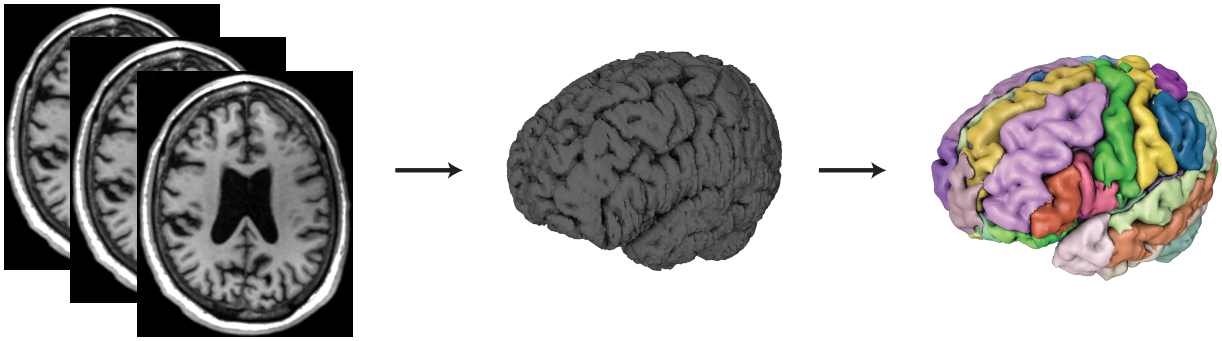
Many other segmentation tasks are performed across medical specialities, and several of these tasks share the characteristics of the examples mentioned above in being time-consuming and error-prone. Moreover, the cost of healthcare systems is increasing globally, both in absolute numbers and relative to GDP. The World Health Organization's (WHO) Global Health Expenditure Database (GHED) index has increased from 8.63 % of GDP to 9.83 % of GDP between 2000 and 2019, which has exceptionally risen further in recent years due to the COVID-19 pandemic, see `https://apps.who.int/nha/database` and WHO et al. (2022). Consequently, a growing potential exists for (semi-) automatic computer models to release doctors from repetitive and time-consuming manual tasks such as segmentation while improving diagnostic precision. These models can also make expert knowledge available for people with limited access to well-trained medical doctors.

1.2 Automating segmentation tasks

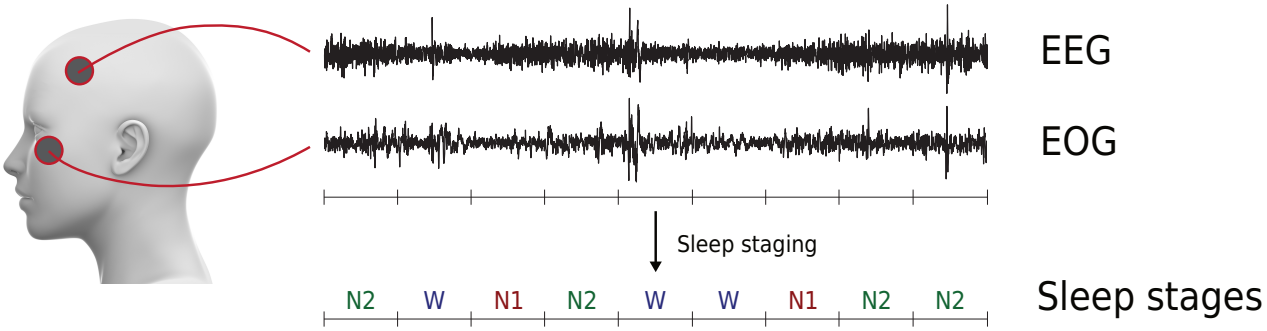
The development of automatic and computer-assisted medical segmentation models has been studied for over five decades (Pal et al. 1993), aiming to improve diagnostic throughput and to minimize the inter- and intra-rater variability of manual segmentation. The technical foundations and historical development of automatic segmentation models can be found in the Background Material chapter 3. At its core, automatic segmentation involves determining a mathematical function that maps input data, such as a medical image, to a corresponding segmentation mask. This mask establishes



(a) Image slices extracted from different 3D medical image volumes showcase several MRI and CT segmentation tasks. Publicly available and anonymized data from the Medical Segmentation Decathlon (`http://medicaldecathlon.com/`, Simpson et al. 2019), Osteoarthritis Initiative (OAI, `https://www.niams.nih.gov/grants-funding/funded-research/osteoarthritis-initiative`) and the 2012 MICCAI Multi-Atlas Challenge (Landman et al. 2012).



(b) Illustrative example of the image quantifications made possible by segmentation. A stack of 2D MRI slices (left image) is first segmented to separate the brain from the skull and other background areas (centre image) and then further into distinct parts of the brain (right image).



(c) Illustration of the sleep staging segmentation process. EEG and electrooculography (EOG) times series data is recorded from a sleeping subject and segmented into 30-second block sleep stages. W is the Wake stage, N1 the Non-REM stage 1, and N2 the Non-REM stage 2 (see Background Material, section 3.3).

Figure 1.1: Image and time series segmentation examples and illustrations.

a relationship between each pixel in the image and a predefined set of regions, also called labels or classes. For example, areas of interest may include one or more organs or lesions.

Automatic segmentation models can serve multiple purposes if found to demonstrate reliability and accuracy across the relevant clinical use cases. It can replace a human rater (computer automation), offer additional information to human raters (computer assistance) or be integrated within a group of human raters to augment its capabilities or substitute a human member. The rapid and consistent segmentation of automatic scoring presents a significant potential increase in diagnostic throughput. A notable example is in population-based mammographic screenings, where two independent human raters typically evaluate each mammogram. An alternative reading protocol can be introduced by replacing one human rater with an automatic scorer and requiring the second human rater's input only if there is a disagreement between the human and automated rater. This would effectively reduce the manual scoring workload by up to 50 % while still ensuring that a human rater reviews all samples (Lauritzen et al. 2022).

The development of automatic segmentation models, however, is non-trivial. The complex nature of medical data, which may vary significantly between patients and recording equipment, makes it challenging to design reliable and accurate segmentation models for many tasks (de Bruijne 2016). The manual programming of a set of rules for segmenting, e.g., a brain tumour, is a complex and often an infeasible task, although early such attempts pioneered the field several decades ago (see Background Materials chapter 3). In recent years, however, the field of automatic segmentation has experienced a significant transformation, driven by increased access to massively parallel computing infrastructure, a growing pool of available data, and algorithmic advancements (Yoshua Bengio, Lamblin, et al. 2006; Hinton et al. 2006; LeCun, Bottou, et al. 1998; Rumelhart et al. 1986). These factors have facilitated the automatic learning of complex segmentation functions with machine learning (Abu-Mostafa et al. 2012). Artificial neural networks from the sub-field of deep learning have been particularly successful (Yoshua Bengio, Courville, et al. 2013; Goodfellow et al. 2016). As detailed in Chapter 3, these methods allow segmentation functions to be learned from observed examples of input-output mappings, such as images or time series to their corresponding segmentations, without the need to manually program the underlying logic for performing the segmentation task.

The 2022 edition of the Medical Image Computing and Computer Assisted Intervention (MICCAI) conference exemplifies this recent activity in the field. This one conference featured presentations of 574 full papers and 38 competitive *challenges*, where teams compete to develop (often machine learning-based) models for specific medical image analysis tasks (L. Wang et al. 2022). Another prominent illustration is the highly successful *U-Net* model, a deep learning model designed for biomedical image segmentation (detailed in Chapter 3.2). Since its introduction in 2015, the U-Net has been applied to many medical image segmentation tasks. The original U-Net papers have been

cited over 58,000 times, as indexed by Google Scholar, March 2023 (Falk et al. 2019; Ronneberger et al. 2015).

1.3 From research to clinical adaptation

Despite the research activity in the field, however, relatively few machine learning models have become commercialized, fewer received clinical approval, and only a handful has seen widespread clinical adaptation. As of October 2022, a total of 521 "AI/ML Enabled Medical Devices" were approved by the U.S. Food and Drug Administration (FDA) for the U.S. market, according to an analysis of the FDA itself[1] based on publicly available information from sources such as The American College of Radiology Data Science Institute (ACR DSI, `https://aicentral.acrdsi.org/`). While arguably a sizable number, this number includes devices for a wide range of tasks (mostly in radiology), including data acquisition, preprocessing, management and visualization, various detection, segmentation and diagnosis tasks and patient triaging and prioritization. Many implement machine learning solutions only as additional features supporting other main purposes. Only a minor subset of approved medical devices perform image- and time-series segmentation, and most have not received complete adaptation by the corresponding medical community. Noticeable exceptions include computer assistant detection (CADe) models for mammography screenings and lung nodule detection in X-ray and CT, which were first approved by the FDA already in 1998, 2001 and 2004, respectively, and for which similar products are routinely used in many hospitals today (Giger et al. 2008). Most other medical image segmentation tasks are still performed manually.

Multiple medical devices have also been approved for tasks related to medical time series data, e.g., PSG sleep studies. One of the very first FDA-approved "AI/ML Enabled Medical Devices" was the Compumedics Sleep Monitoring model by Computmedics Sleep Pty. Ltd., which received approval in 1997 (510(k) Premarket Notification K955841), which was first and foremost a PSG data collection and visualization software tool, but also included early attempts at automated scoring of sleep stages, respiratory events and arousals (although, these models were, arguably, primarily rule-based rather than machine-learning-based at the time). While this and many later models have been extensively used – and still are – for data recording, management and visualization features, the scoring of sleep events, including segmentation of sleep stages, is still performed primarily manually.

So what is missing for more automatic models to be adapted in clinical practice? This is a broad issue with complications in all process steps, from research to adaptation, including many legislative, economic, technical and ethical barriers, some of which are discussed below. However, a central postulate of this thesis is that the main limitation for adaptation of most former and current

[1] `https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices`

automatic segmentation models is the insufficient ability to – or lack of proven ability to – perform robustly under significant clinical variability.

Machine learning models are generally trained and evaluated by collecting and annotating a (preferably large) dataset. A subset is used to train the model, and the remaining to assess the model. According to standard learning theory, with some exceptions, such as in online learning applications, to accurately assess a model's performance on new data not seen during training, the training, evaluation, and future data must be sampled from the same data-generating distribution (Abu-Mostafa et al. 2012).

While all modern machine learning models for medical segmentation follow basic principles of training and evaluating models on separate splits of a dataset, it is often neglected to study or discuss the effect that the developed model may not work well in clinical practice, where the observed data may systematically differ from the data used to train and evaluate the model. It is a fundamental complexity in machine learning for healthcare that data generated in one clinical site (e.g., where the training data was collected) may not represent the whole distribution of clinical data, as other scanners or recording equipment may be used and with different settings on patient cohorts of different demographical- and disease backgrounds.

Machine learning models, including modern deep learning models, are notoriously sensitive to even minor shifts in data distributions between the data to which the model was trained and the data on which the model is later applied. Consequently, a machine learning model trained on even large datasets collected from one or a few different clinical sites and cohorts cannot necessarily be expected to generalize to data from patients in other clinical sites (D'Amour et al. 2020; de Bruijne 2016; Kaushal et al. 2020). A clinically robust model, one that generalizes to most of the expected clinical variability, must either be locally (and preferably easily) adaptable to data from the specific clinical- site and cohort of interest, e.g., by training a new instance of the model or by employing fine-tuning of the model or similar so-called transfer learning techniques (see Opbroek et al. (2015) for an example, and Tan et al. (2018) and Bozinovski (2020) for reviews). Alternatively, the model can be trained on large, cross-cohort, cross-clinical site datasets representing as much clinical variability as possible.

In 2018, the United Nations (UN) agencies of the World Health Organization (WHO) and the Telecommunication Union (ITU) suggested a benchmarking process for evaluating ML/AI models on external, confidential testing datasets where, quote, Wiegand et al. (2019): *"Ideally, test data will originate from various sources to determine whether the use of an AI model can be generalized across different populations, measurement devices, and healthcare settings"*. However, many recent medical devices do not prove general clinical generalization ability. The commentary article of Wu et al. (2021) investigated 130 recently FDA-approved AI/ML-enabled medical devices and identified

several limitations in the evaluation studies on which the approval was based. First, 93 devices did not publically report the number of geographically distinct clinical sites used to evaluate the medical device. Although the FDA has this information, it limits the clinical end user's ability to judge the clinical resilience of the device. Four of the 41 remaining devices were evaluated on only one geographical site and eight on only two sites. This is often insufficient because many factors may vary depending on geographical location, including demographic characteristics, the physical imaging or recording equipment used to collect data, the preprocessing applied to the recorded data and more. Secondly, 126 out of 130 devices were evaluated only on retrospective data, which further limits our ability to study how the devices work on new clinical data and whether the model has other unintentional effects on the practical clinical process. For instance, if clinicians unintentionally use the model in ways it was not designed for or if the model biases the clinician's opinion in an unintended manner (Parikh et al. 2019).

The inability to prove if a machine learning model, whether presented in a research paper or a clinically approved product, can perform well in clinical practice despite a high degree of data variability raises a significant issue, which may limit the general trust and speed of adaptation of machine learning models in healthcare. This thesis developed new optimization strategies and machine learning models that emphasize clinical generalizability to advance the clinical adaptation of machine learning models to segment medical images and time series. It is, however, essential to note that many other complex boundaries and hindrances also limit the successful adaptation of medical machine learning models in clinical practice. Other equally significant challenges include the following (see also Challen et al. (2019), and de Bruijne (2016) for reviews):

- Machine learning models and deep learning models, in particular, are often called black box methods because they are difficult to interpret. For many medical tasks, it is a strict requirement that the end-user of the model, often a medical doctor, can explain to the patient why a certain decision or diagnosis was made based on the recorded data. For this reason, most ML/AI Enabled Medical Devices today are computer-assistance tools, which do not attempt to replace the judgement of the medical doctors but rather provide additional information and second opinions. For a perspective on these problems and possible solutions, see, for instance, Rudin (2019).

- Complex bias issues, which include both unintentional model bias in which, for instance, the developed model underperforms on specific patient demographics, but also the often unclear effect that the usage of the machine learning model has on medical personnel in clinical practice, and how this affects the quality of the diagnostic process (Char et al. 2018).

- Machine learning models applied to medical tasks require a low margin for error, often requiring

levels of accuracy that match or exceed those of human experts performing the same task.

These and other problems were considered in parts of this thesis but were secondary to the main focus on clinical robustness.

Chapter 2

Objectives

The overarching goal of this thesis was to develop clinically robust automatic segmentation models for medical images and time series based on recent advances in machine learning. The main objectives were the following:

1. To develop a task- and scanner-generalizable machine learning pipeline for 3D medical image segmentation problems that requires minimal manual hyperparameter-tuning when transferred to new tasks and are statistically- and computationally efficient.

2. To investigate the applicability of fully convolutional neural network models for time series segmentation problems such as sleep staging, focusing on generalizability across patient cohorts.

3. To develop a clinically robust machine learning model for accurate sleep staging across patient populations, recording hardware and preprocessing pipelines.

2.1 Defining clinically robust machine learning models

In the context of this thesis, clinical robustness is loosely defined as the ability of a machine learning model to perform well across variable data collected in clinical practice, *or*, depending on the context, a machine learning pipeline (here defined as a model, an optimization strategy used to train this model, and all hyperparameters controlling the model and optimization) that is hyperparameter-stable. Hyperparameter-stable means the ability of a machine learning pipeline to be applied (i.e., a new model instance trained to a new set of training data and evaluated on a new set of evaluation data) to a new task without requiring manual hyperparameter tuning while retaining high performance. Here, performance is measured using an overlap metric, such as the F1/Dice score, between the segmentations output by the machine learning model and manually defined expert annotations on a hold-out dataset not used for training the model. When available, the model and individual human experts are compared to a consensus segmentation of a group of human experts to evaluate the relative performance of the model to human scorers. An absolute performance level regarded as sufficient for clinical use is not generally defined, as it will vary depending on the specific application. However, as a rule of thumb, the goal is for a machine learning model to be at least as accurate as individual human experts compared to consensus scores.

This thesis explored two different approaches to obtaining clinically robust machine learning models corresponding to the two definitions of robustness outlined above:

- In Part II (Medical image segmentation), hyperparameter-stable machine learning pipelines were investigated for medical image segmentation tasks. While a specific instance of a machine learning model produced by such pipeline for a single task may not itself be transferable with high performance to new tasks, imaging modalities or patient populations, the stability of the training pipeline itself would allow (also non-technical) end-users to run the pipeline on their specific datasets to obtain a model suitable for a given task and to periodically re-train the model to adapt to continuous data-drift.

- In part III (Sleep staging), it was investigated if a single instance of a model (i.e., the output of a machine learning pipeline) for the task of clinical sleep staging could become robust to a wide range of clinical variability by training the model on large and heterogeneous datasets collected from many sources. By simultaneous training on data from various sources and cohorts without extensive attempts to standardize the data, the goal was to establish an optimization problem in which the only feasible solution did not rely on features specific to individual data recording equipment, patient demographics and more, and which would result in a single model applicable to a wide range of clinical data without requiring on-site re-training.

If feasible, the latter approach is preferred but requires access to large and heterogeneous datasets, which may only sometimes be available, in which case the method studied in Part II may be preferred. In addition, we hypothesized that a machine learning pipeline that performs well across tasks is likely to perform better under the second scenario of simultaneous training on highly heterogeneous datasets. The two approaches should, therefore, be studied in conjunction. See Part IV for further discussions.

Chapter 3

Background material

3.1 Automatic segmentation

An automatic segmentation model is a mathematical function that receives input data on a grid (for instance, a medical image or volume or a physiological time series recording) and outputs a segmentation mask. The segmentation mask specifies an integer value representing a segment or class relationship for each grid position of the input (e.g., each pixel of an image or each time point of a time series) so that all grid positions that share the same segmentation mask value belong to the same semantic class. For instance, all pixels in an image with segmentation mask values of 0 may belong to a background class, while others with mask values of 1 and 2 belong to an organ and tumour, respectively. In this thesis, we consider input data sampled on regular grids, typical of image data and time series where each time step is sampled at a constant rate, and segmentation models that are mappings of the form $f_\theta : X \to Y$ parameterized by θ, where $X \in \mathbb{R}^{d_1 \times \cdots \times C}$ and $Y \in \mathbb{N}^{d_1 \times \cdots \times K}$ with d_1, d_2, ... being one or more spatial dimensions or a single temporal dimension, C the number of channels, and K the number of distinct segmentation classes. Examples of data with multiple channels are multi-modal MRI, where the same target is imaged using multiple different sequences (a group of settings controlling the MRI acquisition) or multi-channel EEG, where brain activity is recorded from potentially several distinct physical EEG electrode positions simultaneously (see Background Material section 8.2 below for details).

The segmentation function f_θ can generally represent a set of logical rules (often manually programmed), a statistical model learned from observed examples of data mappings from an input to a desired output, or some combination of the two, e.g., a rule-based system where some parameters controlling the rules are learned from data, or where a statistical model is used to refine or improve the outputs of a rule-based system. Taking image segmentation as an example, early methods (approximately 1970 to late 1990s) predominantly applied logical rules by processing images first at the pixel level by applying intensity thresholds, edge-detection filters and region-growing algorithms, on top of which pre-defined mathematical models of specific shapes of interest could be fitted to detect, for instance, an ellipsoidal structure of interest. For a review, see, e.g., Pham et al. (2000). These methods were naturally limited in detecting complex and variable structures, which may be difficult to model mathematically. In the late 1990s, as larger quantities of digitalized medical data became available, methods that rely on a library of examples were developed, e.g., active shape models

(Cootes, Edwards, et al. 2001; Cootes, Taylor, et al. 1995) and atlas registration methods (Maintz et al. 1998).

In statistical learning or machine learning, one seeks to discover or learn a function which solves a task of interest on a set of (often labelled) training examples while also being able to predict accurately new examples not seen before (Abu-Mostafa et al. 2012). Because machine learning allows the generation of predictive models without requiring explicit programming of task-specific rules, it became possible to develop, for instance, medical image segmentation models for complex tasks much more rapidly. However, to successfully learn functions on image data, it was necessary to compute (typically manually defined) features of the images, which summarize various aspects of the image into a lower-dimensional vector, which could be used for subsequent tasks such as classification or segmentation (Remeseiro et al. 2019). For instance, Sobel filters were often used to approximate the gradient of the image function and detect edges (Sobel et al. 1973), while Gabor filters were used to detect textures (Gabor 1946). The resulting features could then be clustered, and a histogram constructed by counting the occurrence of each feature in local regions (patches) of the image.

Combining feature extraction and machine learning led to many successful medical image analysis applications, such as computer-aided models for breast cancer diagnosis (Suri et al. 2006; Wernick et al. 2010). However, the manual definition of features may limit the ability to transfer such models to other tasks or patient populations, and the model's performance is ultimately limited by the designer's ability to define discriminative yet low-dimensional features for a given task (Litjens et al. 2017). For these reasons, the ability to combine machine learning models with automatic feature learning (also called representation learning) has been extensively studied (Yoshua Bengio, Courville, et al. 2013). Since around 2000, deep learning, a sub-field of machine learning focusing on deep neural network models that apply composites of simple feature extraction functions that together learn progressively more complex and abstract features (Goodfellow et al. 2016), has been particularly successful. This success can be attributed to more widespread access to massively parallel computing infrastructure, increasing available training data, and algorithmic advances (Yoshua Bengio, Lamblin, et al. 2006; Hinton et al. 2006; LeCun, Bottou, et al. 1998; Rumelhart et al. 1986). Deep learning has seen several successful applications across domains, including medical image and time series processing (Faust, Hagiwara, et al. 2018; Hesamian et al. 2019; Razzak et al. 2018). Deep learning contrasts the manual design of filters for feature extraction by enabling the automatic learning of (compositions of) convolution filters tailored to a specific task. With sufficient training data and proper tuning of the model's approximation capacity to the task at hand, it is thus possible to learn deep models that approximate complex segmentation functions directly on, for instance, image inputs, sidestepping the need for manual feature engineering.

However, deep learning has yet to trivialize the development of automatic medical segmentation

models. Deep learning models may be developed without relying on task-specific expert knowledge but still benefit significantly from such being utilized in the design and training of a model. Moreover, determining which task-specific knowledge can be omitted when using deep learning often necessitates expertise in deep learning models and their optimization instead. For instance, the inductive bias (a set of imposed underlying assumptions about the unknown target function) of different deep learning architectures may make them suitable – or not – to learn various problems. Knowledge of the segmentation problem and data may also reveal helpful in- or equivariance properties that a model may benefit from, e.g., equivariance to translations or rotations of an object to be segmented within a large image volume. Expert knowledge may also be necessary to construct loss functions and evaluation metrics that guide the learning and model selection process to solve a task more effectively. In general, deep learning models are controlled by a usually extensive set of hyperparameters defining both the model architecture and optimization, which often requires manual tuning for the resulting deep learning model to function well.

This thesis explored deep learning architectures and in- and equivariance properties to data transformations that support efficient learning of clinical robust segmentation models for medical images and time series to increase model performance and reduce the need for expert knowledge on deep learning when implementing such systems in clinical practice.

3.2 Medical image segmentation

3.2.1 Convolutional neural networks

A specific deep neural network architecture, the convolutional neural network (CNN), has been particularly influential for image processing tasks since its invention in 1980 and later when found able to learn the automatic processing of hand-written digits with previously unseen accuracy (Fukushima 1980; Lecun et al. 1998). Since then, most medical image and time series processing systems based on machine learning have relied on CNNs. CNNs have been successfully applied to tasks as diverse as mammography screenings (Hamidinekoo et al. 2018), brain MRI segmentation (Akkus et al. 2017), liver tumour segmentation (Bilic et al. 2023), ultrasound analysis (S. Liu et al. 2019), classification of pulmonary tuberculosis (Lakhani et al. 2017), lung nodule segmentation (S. Wang et al. 2017), positron emission tomography (PET), CT, MRI reconstruction (Lundervold et al. 2019; Reader et al. 2020; Würfl et al. 2016) and several others.

CNNs are highly effective for processing images because they model complex structures in images as composites of simpler, more local features, similar to how the human visual cortex operates (Hubel et al. 1968). This is due to their unique architecture, which has multiple inherent inductive biases useful for the effective processing of data on regular grids such as images:

1. Local connectivity & parameter sharing: Unlike classical, fully connected neural networks that implement complete matrix multiplication, CNNs use convolutional layers to implement local transformations with shared weights. This enables the detection of local patterns in a translationally equivariant manner. These biases build on the assumptions that spatially close pixels are more relevant to each other than distant ones and that similar shapes, textures, or objects can appear in multiple locations within an image and may be arbitrarily translated.

2. Translational invariance: CNNs typically implement pooling layers in addition to convolutional layers. Pooling layers reduce the dimensionality of the feature maps by aggregating local features within a pooling region. Typical pooling operations include max-pooling, where the maximum value in each pooling region is taken as the output. Pooling introduces translation invariance to minor translations (smaller than the pooling window).

3. Hierarchical feature learning: CNNs are deep compositions of convolutions and (usually) max-pooling layers which learn progressively more abstract features. The first layer may compute simple approximations of the first derivative of the image function to detect edges, which in subsequent layers combine to detect local edges and textures. The convolutions applied in deep layers may ultimately respond to complex patterns in the input image, such as a tumour.

In combination, the inductive biases of CNN architectures introduce global translation equivariance (i.e., translation of an object in the image over longer distances will result in correspondingly translated feature maps) and invariance towards more local translation (i.e., feature maps do not change if the translation is small). The use of convolution operations inherently models the locality of pixel data in a parameter-efficient way. For further details on CNNs and their historical development, see reviews by Schmidhuber (2015) and Gu et al. (2018).

3.2.2 Fully convolutional neural networks

Classical CNN architectures were designed to process input data on a regular grid, such as 2D images or 3D volumes, and produce a single output. Several convolution and max-pooling operations are applied to the input image to extract a stack of features, which are then flattened and fed into a fully connected layer for regression or classification. To use CNNs for semantic segmentation, where all points on the input grid need to be assigned a prediction, most CNNs have historically been applied to overlapping sub-grids of the input, also known as patches. See Guo et al. (2019) and Prasoon et al. (2013) for examples. This approach conveniently solves two practical problems: First, by training on patches, the CNN is exposed to many more unique training examples, as more patches can be generated from a single image. Early research considered this necessary because

most medical applications had small training datasets. Secondly, the use of patches addresses the polynomially expanding computational and statistical complexity of segmenting larger images, as the size of patches can be kept constant even for large images. This was also considered necessary, as computational resources, particularly memory on graphics processing units (GPUs), limited the size of images that could be processed in a single forward pass through the CNN model. The downside to patch-wise processing is that it restricts the ability of the CNN model to process long-range dependencies, which could hinder the detection of structures with low local contrast and may lead to false-positive predictions for pixels far away from the object to be segmented. In addition, the model has to be applied once for each pixel to segment, which is computationally expensive and wasteful, as many of the convolutional computations are repeated with each forward pass.

Since 2015, fully convolutional networks (FCNs) have become popular models for segmentation tasks (Long et al. 2014). FCNs are encoder-decoder networks that initially encode an image through several convolutional and pooling operations to extract progressively more abstract features of lower spatial dimensionality, like CNNs. However, unlike CNNs, FCNs do not flatten the extracted feature maps to produce a single classification. Instead, they perform a decoding step that applies a learned up-sampling function to project the feature maps back to the input dimensionality to obtain a semantic segmentation in a single forward pass. A schematic example can be seen in Figure 8.1 of Paper C and the U-Net papers of Falk et al. (2019) and Ronneberger et al. (2015).

The decoder sub-network applies composites of learned up-sampling functions, typically transposed convolutions, nearest neighbour, or bilinear upsampling, followed by standard convolution operations. However, because the pooling operations of the encoder sub-network discard some spatial information, the direct learning of the up-sampling function is ill-defined. To overcome this, networks like U-Net implement skip connections (formulated early, see, e.g., Bishop et al. 1995; popularized by the ResNet paper of K. He et al. 2016a). By passing feature maps from various levels of the encoder network to the decoder, which are concatenated with, added to, or otherwise combined with the decoded feature maps, spatial information may be recovered (Drozdzal et al. 2016). Skip-connections also enable the automatic learning of what scale the image should be processed because the relative weight of information passed from skip-connections and features from deeper layers can be adjusted through the learning process. Finally, skip-connections have been empirically found to support easier optimization of deep neural networks, which otherwise train slowly and tend to display accuracy degradation (even on the training set) as large numbers of layers are stacked (K. He et al. 2016a; K. He et al. 2016b; Szegedy et al. 2017).

FCNs are highly effective at image segmentation due to their computational efficiency and ability to process the entire input image or larger patches in a single pass (up to computational resource limitations, see below), providing a more comprehensive context. This advantage often allows FCNs

to outperform standard CNNs. In addition, perhaps surprisingly, despite typical FCNs having millions of parameters (e.g., around 30 in the default U-Net implementation), FCNs exhibit relatively high statistical efficiency. The number of required training images varies depending on factors such as the segmentation task's complexity, image variability, and the quality of manual segmentations. For relatively simple tasks, like segmenting organs without lesions in MRI or CT volumes, U-Net-like models can learn accurate segmentation functions using only 20-40 annotated image volumes (see, for instance, the cardiac and spleen segmentation tasks of Simpson et al. 2019). However, more complex segmentation tasks involving lesions or other variable targets may require hundreds or thousands of examples to achieve robust performance.

FCNs were a primary focus of this thesis because of their proven ability to work well across diverse tasks in medical images. It is, however, essential to note that concurrently to this thesis, many other strong candidate models were developed, including various extensions to FCNs and the U-Net, but also based on a more recent deep learning architecture called the Transformer (Vaswani et al. 2017). The Transformer has significantly advanced the field of natural language processing and has seen several adaptations to images (natural as well as medical, see, e.g., Dosovitskiy et al. 2020; Ze Liu et al. 2021) and time series data (J. Li et al. 2023; Phan, Mikkelsen, et al. 2022). Future research directions involving a combination of FCN and Transformer architectures are discussed in Part IV chapter 15.

3.2.3 Efficient segmentation of 3D images

Most FCNs, like the U-Net, were initially designed for 2D image segmentation problems. They are easily generalizable to N dimensional data by implementing N dimensional convolution, max-pooling, and up-sampling operations. See, for instance, our Paper C in Part III on U-Nets for 1D time series. Medical images are often volumetric ($N = 3$ dimensions) and can thus be segmented using, e.g., a 3D U-Net (Çiçek et al. 2016). However, processing large 3D volumes is computationally expensive, and significant GPU memory is required, particularly during training, where the outputs of intermediate layers must be stored for referencing during backpropagation. Even on modern GPUs, training 3D FCNs usually requires small batch sizes, downsampled images or reduced model size (e.g., fewer layers, filters per payer or kernel sizes) to reduce memory consumption, each of which may be sub-optimal depending on the task. Another approach is to use a 3D FCN model applied to smaller 3D subsets of the volume (volumetric patches). However, this method inherits some of the limitations of patch-based segmentation with CNNs discussed above.

3D FCNs have achieved numerous successful applications despite their limitations (Çiçek et al. 2016; W. Li et al. 2017; Milletari et al. 2016), but another popular and viable approach involves segmenting 3D images using 2D models (P. F. Christ et al. 2016; Norman et al. 2018). This method

reduces computational overhead while maintaining the statistical efficiency of 2D kernels. For example, a 2D model can be applied to each 2D slice along one of the three orthogonal image axes separately to construct a complete 3D segmentation volume. However, 2D slicing is still a patch-based method, and although each 2D image displays the entire context along two axes, it provides no information along the third axis. Various efforts have been made to combine the statistical and computational efficiency of 2D segmentation models with better utilization of 3D image information. Popular approaches include using models that observe a small number of 2D slices along the third axis (sometimes referred to as 2.5D models) (Xia et al. 2020), training multiple 2D models that process slices from a distinct image axis and ensemble their outputs, training a single model that incorporates information from multiple planes simultaneously (H. R. Roth et al. 2016), and employing cascaded model setups where one model generates an initial segmentation based on 2D or downsampled 3D inputs, which is then refined by a 3D model operating on smaller (full resolution) patches (Isensee et al. 2018; Jiang et al. 2020; H. Roth et al. 2018). In Part II Paper A, we extend these ideas by training an efficient 2D FCN model on many randomly oriented image planes simultaneously and utilizing the imposed rotational equivariance to output ensemble predictions made over multiple view orientations.

3.2.4 Clinical generalization across cohorts, scanners and tasks

The introduction of CNNs, particularly FCNs, has allowed significantly easier and faster development of segmentation systems for new medical image segmentation tasks. In contrast to earlier methods based on programmatic rules, mathematical modelling or manual feature engineering (outlined in Chapter 1), deep learning segmentation models require little task-specific knowledge to develop. They are quickly trained on modern massively-parallel hardware. Hundreds of deep learning models that solve particular segmentation tasks in medical scans acquired on similar hardware and from a demographically and geographically narrow cohort of people are published yearly. However, as outlined in Chapter 1, it is often not clear if the developed models transfer to images from other scanning equipment or different patient cohorts or how the machine learning pipeline (the model and its optimization routine) transfers to other, but related tasks such as a separate organ or lesion type. These limitations reduce the clinical adaptation of automatic segmentation systems.

In 2018, the MICCAI conference hosted the Medical Segmentation Decathlon (MSD) challenge (Simpson et al. 2019). The MSD challenge aimed to identify models and optimization techniques that support the development of clinically robust machine learning models that can automatically transfer to different segmentation tasks. The aim of the challenge thus was to encourage the development of segmentation models that adhere to the first of this thesis' two definitions of clinical robustness outlined in section 2.1, in that each model instance does not necessarily need to be robust across

tasks or cohorts as-is, but the machine learning pipeline that produces it should. The pipeline is then easily transferable and also applicable in clinical practice where experts on the manual tuning of machine learning pipelines may not be available.

Teams were invited to develop a machine learning pipeline which could be applied to a new segmentation task without human intervention and by receiving only a set of labelled training examples as input. No task-specific information was supplied with each task, and manual tuning of, for instance, model architecture or optimization hyperparameters was not allowed. The participating pipelines should generalize as well as possible across ten distinct medical image segmentation tasks; 7 were known to the developers during the first phase of the challenge, and 3 were revealed only in the second phase after the development systems were locked and evaluated.

No restrictions were imposed on which techniques could be used to obtain a clinically robust pipeline. Paper A of Part II of this thesis introduces our team's contribution to the challenge, the Multi-Planar U-Net (MPUnet) model, described and summarized below. Other groups took different approaches, the most noticeable of which are described and discussed in the Related Work section 7 of Part II below where the findings of the 2018 MSD challenge paper (Simpson et al. 2019), which we co-authored, are summarized.

3.3 Sleep staging

3.3.1 The physiology of sleep

Sleep is essential to human health, and abnormal sleep is associated with a wide range of morbidities, including psychiatric-, neurological- and cardiovascular diseases and stroke (Baandrup et al. 2018; Chattu et al. 2018; Ponsaing et al. 2017). It is estimated that upwards of 20 % of the Danish population suffer from sleep disorders such as sleep-disordered breathing (SDB) (P. Jennum et al. 2009). An epidemiological study in the Netherlands reported a prevalence of 5.3 % to 12.2 % for each of 6 major sleep disorders (insomnia, circadian rhythm sleep disorders, parasomnia, hypersomnolence, restless legs syndrome and SDB, respectively. See Walker et al. (1990), chapter 77, and Sateia (2014) for definitions) and an overall prevalence of sleep disturbances of 32.1 % Kerkhof (2017). Consequently, sleep disorders constitute one of the most common diseases and impose significant individual and societal costs (Garbarino et al. 2016).

The *sleep cycle* is central to sleep physiology and differentiating healthy from abnormal sleep. During normal sleep, the brain and body transition through multiple distinct physiological phases called *sleep stages*. The current standard for sleep scoring defines five different stages of sleep; Wake (W), non-REM stages 1, 2 and 3 (N1-N3), and rapid eye movement (REM) sleep (Iber et al. 2007). The sub-classification of non-REM sleep into N1–N3 signifies deeper levels of non-REM sleep. These

are described in more detail below. Each stage is associated with specific physiological functions, and during normal sleep, the brain transitions between non-REM and REM sleep in characteristic cycles of approximately 90 minutes. While significant variation occurs, the canonical sleep stage transition is Wake → N1 → N2 → N3 → N2 → REM with subsequent cycling between non-REM and REM stage sleep with increasing duration of REM with each 90 minutes cycle. About 75 % of healthy sleep is spent in non-REM sleep, with N2 the single most common stage occupying approximately 45 % of the night's sleep (Feinberg et al. 1979; Patel et al. 2022). The durations, however, vary, with older people typically getting less N3 sleep and more N2 sleep.

3.3.2 Polysomnography & manual sleep staging

Sleep stages can be objectively detected with reasonable certainty because each stage is associated with particular physiological activity, which can be picked up using external recording equipment. The gold-standard method for objectively detecting sleep stages and other sleep parameters is *polysomnography* (PSG). PSG is a sleep study involving the continuous overnight recording of multiple brain and body signals. See Supplementary Figure C.1 of Paper C for an example. These typically include neurological activity measured using EEG, eye movement measured using electrooculography (EOG), muscle activity measured using electromyography (EMG), heart rate measured using electrocardiography (ECG) and a range of non-bio-electrical signals such as body temperature and position. Each variable presents distinct patterns in each stage of sleep, which can, therefore, be detected and mapped throughout the night.

The number of sleep stages, their corresponding physiological basis, and the rules used to score them have varied. Rechtschaffen and Kales (R&K) proposed the first standard for sleep stage scoring in 1968 (Kales et al. 1968). Today, a simplified set of stage scoring rules by the American Academy of Sleep Medicine (AASM) guidelines are used almost universally (Iber et al. 2007). The standard defines a set of rules used to score each stage. Stage Wake, which represents the waking state from full alertness to drowsiness, is primarily determined by the AASM guidelines for adults by the occurrence of an 8 – 13 Hz sinusoidal activity (the so-called alpha rhythm) in occipital EEG electrodes (see technical specifications below) when eyes are closed, REMs when eyes are open, and eye blinks. Stage REM is defined by the occurrence of REMs when the eyes are closed, low baseline EMG (muscle) tone with transient bursts of EMG activity and so-called sawtooth 2–6 Hz visually sharp or triangular EEG waves. The non-REM stages represent a continuum of decreasing brain activity and physiological arousal: stage N1 represents the transition from wakefulness to sleep, marked by a decrease in alpha rhythm frequency (8 – 13 Hz) and an increase in theta activity (4 – 7 Hz) in the EEG; stage N2 is characterized by the presence of sleep spindles (short bursts of 11-16 Hz activity) and K-complexes (sharp, high-amplitude negative peak followed by a slower positive component) in

the EEG; and stage N3, also called slow-wave sleep, is characterized by the presence of delta waves (0.5 – 2 Hz) in the EEG. During non-REM stages, heart rate, blood pressure, and respiration rate generally decrease, and rapid eye movements are absent. See Supplementary Table C.1 of Paper C for a brief overview of the characteristics of all five stages. Note that variations to the AASM guideline are used when scoring sleep in infants (0 – 2 months) and children (from 2 months and with no strict upper age boundary, see M. Grigg-Damberger et al. 2007).

AASM stages are usually scored in fixed, contiguous segments of 30 seconds. The mapping of a night's sleep into sleep stages is called *sleep staging*. The output of the sleep staging process is a time-indexed graph with stages on the y-axis called a *hypnogram*. In manual sleep staging, trained medical doctors or technicians inspect each 30-second window of PSG data and apply the AASM rules. If the features supporting two or more stages are present within the same scoring block of 30 seconds, the stage which occupies the majority of the segment is scored.

Technical specifications The AASM guidelines also establish a set of recommendations for which EEG and EOG channel derivations should be used to score each stage, as well as various technical and digital specifications such as recommended sampling rate and preprocessing filtering settings. EEG is generally collected by placing electrodes in specific positions on a subject's scalp. Typical positions are defined by the international 10-20 system, in which electrodes are placed at various fractions of distances on lines between easily located skull landmarks, such as on the nasion-inion midline (Jasper 1958). Each site is identified with a letter and a number (for instance, C3), which indicates the area (where C indicates central) and hemisphere (where the odd number 3 indicates placement on the left hemisphere). An EEG amplifier records a voltage differential between an electrode and a common ground electrode, generating a so-called single-ended voltage relative to the ground. However, because the grounding circuit contains noise, it is often necessary to reference the single-ended signals by subtracting another electrode's signal from an active electrode of interest. As the noise of the grounding circuit is similar in all electrodes, this differential EEG derivation will approximately remove the noise from all active electrodes (Luck 2014). In the 10-20 system, the most common reference electrodes are placed on the mastoid bone process behind the left and right ears, known as A1/M1 and A2/M2, respectively, for contralateral referencing.

An EEG montage describes how EEG electrodes are placed and referenced. The AASM guidelines specify suitable montages for visual sleep scoring. For example, they recommend recording at least three EEG derivations, preferably F4-M1, C4-M1, and O2-M1, to cover frontal (F4), central (C4), and occipital (O2) activity. Additionally, it is recommended that the EEG signals be sampled at a minimum of 200 Hz, with a preference for 500 Hz, although the signals can be visually inspected after downsampling. A band-pass filter of 0.3 Hz to 35 Hz should also be applied during preprocessing.

It was, however, a primary focus of our work on automatic sleep staging, introduced in Papers C and D of Part III, that our U-Time and U-Sleep models should be able to function with much less rigorously defined and standardized EEG montages, e.g., where only a single or few EEG electrodes are required, and sample rates and filtering settings may vary between recordings.

3.3.3 Diagnosing sleep disorders

Clinical PSG and the subsequent extraction of sleep stages and other sleep-related events are valuable because they enable a relatively objective analysis of potential disruptions to the normal sleep cycle. As a result, various sleep disorders can be diagnosed based on how they interfere with normal sleep physiology and cycles. Factors that may be affected by a specific disorder include the circadian rhythm, the ability to fall asleep, maintain sleep, and wake up, transitions between sleep stages, the duration of each sleep stage, and the overall development of sleep dynamics throughout the night. For example, narcolepsy, a prevalent hypersomnia disorder, is characterized by excessive daytime sleepiness and abnormal REM sleep patterns. In the case of narcolepsy type 1, the disorder is likely caused by a loss of neurons in the hypothalamus that produce the sleep-cycle regulating neuropeptide orexin. This loss of neurons disrupts the normal sleep-wake balance, leading to the symptoms associated with narcolepsy, which can be objectively diagnosed by scoring sleep stages in a so-called Multiple Sleep Latency Test (MSLT) or "daytime nap study". In an MSLT, the patient takes five regularly scheduled daytime naps in a bed with the lights turned off. Two of the diagnostic criteria for narcolepsy are an average sleep latency of ≤ 8 minutes (the time from lights are turned off to the occurrence of the first non-wake stage of sleep) and at least two so-called sleep-onset REM periods (SO-REMs), which are transitions directly from stage Wake into a REM period (as defined by the AASM International Classification of Sleep Disorders, Sateia 2014).

Other disorders can be detected through characteristic changes in sleep physiology. For instance, people with REM sleep behaviour disorder will move and act out dreams instead of lying still, which may be detected by looking for abnormal body movement during the REM sleep stage. Sleep apnea and other sleep-related breathing disorders, in which prolonged breathing interruptions occur several times per hour, will be visible in the recordings of chest movement, sound (snoring), and blood oxygenation levels.

3.3.4 Limitations of the AASM guidelines and manual scoring

While the AASM guidelines are widely used and have contributed significantly to standardizing sleep medical practice, there are several issues with the current standard for scoring sleep stages. As summarized in the concluding remarks of Silber et al. (2007) (p. 129) in the introductory paper for

the new AASM guidelines for visual scoring developed by the AASM-appointed Visual Scoring Task Force in 2004 – 2005: *"No visual based scoring system will ever be perfect, as all methods are limited by the physiology of the human eye and visual cortex, individual differences in scoring experience, and the ability to detect events viewed using a 30-second epoch"*. Some well-known limitations are:

- Visual scoring of high-frequency 1D signals is not intuitive for most individuals and requires extensive training. While the ability to score accurately according to the AASM guidelines varies with experience, even experienced scorers are subject to significant inter- and intra-scorer variability (Danker-Hopfe, Anderer, et al. 2009; Danker-Hopfe, Kunz, et al. 2004; Drinnan et al. 1998; Rosenberg et al. 2013). Ambiguities and room for interpretation often arise in practice, particularly when recorded data is noisy or otherwise difficult to evaluate.

- Another significant source of uncertainty is the arbitrary segmentation of continuous sleep physiology into precise (usually) 30-second blocks, called epochs. Determining how to assign a single sleep stage to an entire segment when a sleep stage transition occurs mid-epoch can be challenging, and the agreement of human scorers tends to decrease near sleep stage transitions (Kim et al. 1993). The current standard does not support scoring transient sleep stages shorter than 30 seconds. This restriction to scoring in 30-second epochs limits diagnostic precision, as sleep stage durations might be over- or under-estimated, and our ability to study microsleep patterns in the sub-30-second domain may vary between healthy and diseased cohorts (St Kubicki et al. 1996).

- The AASM guidelines primarily describe sleep in healthy individuals, with limited applicability to those with brain disorders, such as Alzheimer's or stroke, which significantly alter typical brain activity and EEG recordings (Finnigan et al. 2013; Ju et al. 2014). This complexity and ambiguity make it challenging to apply standard scoring rules in such cases. Lower inter-rater agreement between expert scorers has been observed when scoring data from patients with Parkinson's disease compared to the overall agreement (Danker-Hopfe, Kunz, et al. 2004).

- Due to the repetitive nature of manual scoring, it is an error-prone process, with the potential for scoring fatigue affecting the results.

- Manual scoring is time-consuming, as discussed in the Motivation chapter 1. This limitation can impact the efficiency and scalability of sleep stage analysis in both research and clinical settings.

Automatic or computer-assisted sleep staging has been suggested to overcome some limitations.

3.3.5 Automatic sleep staging

Attempts to develop tools that assist or automate sleep staging have been made since the beginning of digitalized sleep recordings (Bob Kemp 1993; Penzel et al. 1991; Thomas Penzel et al. 2000). Virtually all work between 1970 and 1990 aimed to implement computational pipelines that carefully mimicked the manual scoring process (the R&K scoring rules, widely accepted at the time Kales et al. 1968). First, the EEG, EOG and EMG recordings were preprocessed to remove artefacts, isolate or suppress specific frequency ranges, and extract features. It was not feasible to directly model sleep stages from the raw, high-frequency recordings due to computational, statistical (e.g., lack of digitalized, annotated data) and algorithmic limitations (e.g., neural networks specialized for long sequence data, such as recurrent- or convolutional neural networks, were not yet invented). The most common feature extraction was based on spectral analysis using the fast-Fourier transform (FFT) algorithm to quantify the frequency components of the recorded signals within each scoring window. Nearly all systems would measure the average amplitude of alpha- and delta waves (8 – 12 Hz & 0.5 – 4 Hz, respectively) and count occurrences of sleep spindles (bursts of 11 – 16 Hz activity during N2 sleep) from the EEG recordings, as well as quantify eye movements and muscle activity in EOG and EMG (Berry et al. 2015; Hasan 1996). Sleep stages would then be inferred from these features using statistical models or rule-based systems designed to follow the R&K guidelines as closely as possible. Several such systems were developed and often evaluated on relatively small (5 – 10 individuals) healthy cohorts. For examples, see Martin et al. (1972), Agnew Jr et al. (1967), Hasan et al. (1993) and Roessler et al. (1970). All methods from this time suffered from the same problem: They were brittle and difficult to apply in practice to new data, as the parameters controlling the preprocessing, feature extraction, and rule-based classification system were largely hand-crafted and required re-tuning for the system to adapt to new signal characteristics, artefacts and recording hardware (Bob Kemp 1993). Together with – at the time – high computational costs and relatively low accuracy, these systems never saw widespread clinical adaptation.

Following the general trend of the development of machine learning outlined in Chapter 1 above, through the 1990s, automatic sleep staging systems increasingly started to implement neural networks (or other classifiers, e.g., support vector machines, random forests, gaussian mixture models or various linear models) to model the mapping from extracted features to sleep stages, thus replacing the fragile, rule-based classification with a function learned from a set of supervised examples (Grözinger et al. 1995; Pfurtscheller et al. 1992; Principe et al. 1989). While this paved the way for promising improvements in scoring accuracy, these early neural-network-based systems still relied on manually defined input features and still suffered from a lack of robustness towards input data variability (Robert et al. 1998). Nonetheless, recent studies have refined and used the strategy of extracting

time- and frequency domain features for subsequent classification using supervised machine learning models. See, for instance, Boostani et al. (2017).

Another prominent example of a successful application of manual feature extraction and subsequent machine learning classification is the recent sleep staging extension to the well-known `YASA` sleep analysis software package (`https://github.com/raphaelvallat/yasa`) (Vallat et al. 2021). This sleep stager was developed concurrently with our U-Sleep model described in Paper D of Part III with a similar focus on clinical robustness. The algorithm first extracts a set of expert-defined time- and frequency domain features in windows of 30 seconds from a single EEG (and optionally EOG and EMG) channel, from which sleep stages are classified using a LightGBM (a gradient boosting decision tree, Ke et al. 2017) model. While U-Sleep was found to perform slightly better than `YASA` on sleep-disordered patients in Vallat et al. (2021), the `YASA` has other benefits, such as being more easily interpretable because of the manual definition of input features as compared to deep learning models like U-Sleep.

Since around 2010, the development of automatic sleep staging models has seen increasing use of recurrent neural networks (RNNs, Rumelhart et al. 1986, and more often their later extension to long short-term memory models, LSTMs, Hochreiter et al. 1997) and CNNs that process raw input signals directly. For a review, see (Fiorillo, Puiatti, et al. 2019). Given sufficient labelled training data, these neural network architectures can learn direct mappings from the raw input signals to sleep stages (although the signals are often preprocessed by re-sampling to lower sample rates and band-pass filtered to isolate relevant frequency bands as recommended by the AASM guidelines). This ability makes the development of automatic sleep stagers significantly more accessible and faster. Learning the scoring function from raw, annotated data examples also allows for discovering better solutions. It is, in theory, possible that a machine learning model learns a much simpler, more robust mapping from EEG signals to sleep stages, which is difficult to detect visually or otherwise unintuitive to humans and, therefore, has not been previously discovered with manually defined feature-based learning or manual scoring. However, modelling sleep stages from raw, high-frequency input signals increases the risk of model overfitting. Larger training datasets are usually required to develop models that operate on the raw input signals compared to manually defined features.

Prominent examples of neural network models applied successfully to raw signals include the work of Supratak et al. (2017), which developed the *DeepSleepNet* model. DeepSleepNet combines both an RNN and CNN into one model. First, a CNN sub-network extracts feature maps from the input signal. These features are then input to a bi-directional LSTM sub-network that outputs the predicted sleep stage sequence. The intention is to let the CNN sub-network extract, in a computational- and statistically efficient manner, time equivariant feature maps that summarize the essential information of the long, high-frequency inputs from which the RNN explicitly models the temporal transitions of

sleep stages. DeepSleepNet scored two different datasets accurately with overall overlap macro F1 scores (the unweighted average of F1 scores computed for each stage individually) of 0.72 – 0.80, a performance likely comparable to human experts, using only single EEG channels as input. These results showed a significant potential to assist or automate sleep staging while simplifying the PSG setup (e.g., for home-testing or wearable devices), which usually involves in-hospital recording using several EEG electrodes. However, the original DeepSleepNet model was individually trained and evaluated on relatively small datasets of 50–100 PSG records and on single EEG channels, making it unclear if the developed models transfer with high performance to other cohort demographics, other EEG channels, or other recording equipment without re-training.

Biswal, Sun, et al. (2018) also proposed a combined RNN and CNN model but interestingly used (in addition to raw EEG in other experiments) spectrogram representations of the EEG signals as input for the initial CNN sub-network. This model was trained and evaluated on two large cohorts of 10,000 and 5,804 PSGs, respectively, and was, likely, the sleep staging model to date trained on the largest and most diverse set of clinical PSG data. Notably, the authors also performed cross-cohort experiments, where their model was trained on one of the two cohorts and evaluated on the other, and one model was trained on both cohorts simultaneously. Such experiments are essential to assess the expected performance in clinical practice and were missing from most previous studies of automatic sleep staging. When trained and evaluated on the largest dataset, their CNN+RNN model had an average Coheen's Kappa (Cohen 1960) coefficient of 76.4 (2 input channels). When training and evaluated on the smaller dataset, the score was 73.4 (2 input channels). When training on the larger dataset and assessing on the smaller, the model scored 73.2, i.e., comparable performance to the model trained and evaluated on the smaller dataset. Oppositely, when trained on the smaller and assessed on the larger dataset, the model scored slightly lower at 69.2, down from 76.4 for the model trained and evaluated on the larger dataset. When simultaneously trained on both datasets, the model scored 74.2 on the smaller and 77.8 on the larger datasets, respectively, thus performing better than each of the two models trained only on the individual cohorts. This critical experiment showed that even if a neural network model is trained on thousands of PSGs (as in the case of the smaller dataset), such a cohort may not contain enough variability to facilitate learning a model that generalizes with similar high-performance to other clinical cohorts. However, pooling heterogeneous datasets from different sources appeared to be a potential future direction for developing clinically robust automatic sleep stagers. These points were recently reiterated by Fiorillo, Monachino, et al. (2023), who found (based on experiments using our U-Sleep model, see Part III, Paper D) that training models on even extensive and heterogeneous datasets (in terms of variability in demographics and diseases represented) but collected from only single clinical site is itself insufficient to ensure solid clinical generalization to other clinical sites.

In Part III, Paper D specifically, although also focussing on the validation of the use of fully convolutional neural network architectures, inducing model invariance to input EEG and EOG channels variability, and the ability to perform high-frequency sleep staging, we built upon the essential ideas of Biswal, Sun, et al. (2018) when developing our *U-Sleep* model, an automatic sleep stager trained on extensive, heterogenous data from 16 independent clinical studies.

3.3.6 Segmentation of time series with CNNs

In this thesis, we consider time series in PSG data where all time steps are regularly sampled, i.e., EEG (and other) physiological signals are sampled or re-sampled at a uniform rate throughout the night. In sleep staging, the goal is to score sleep stages in contiguous, fixed-length segments that span the entire input signal together. Consequently, while often not described as such, sleep staging is a segmentation task as defined in the Background Material section 3.1 above.

Because segmenting a time series on a regular grid is equivalent to segmenting a one-dimensional image, this thesis aimed to modify image segmentation models to the 1D setting to leverage the advances described in the Background Material section 3.2 above and in the enclosed papers of Part II. Specifically, we studied the applicability of FCNs for time series segmentation tasks exemplified by sleep staging. Time series have traditionally been thought to be better suited for processing by sequence models, such as RNNs, when using deep learning techniques. RNNs explicitly model temporal relations and naturally handle variable input and output length data. However, for tasks like sleep staging where the input data is (or can be) sampled on regular grids, and there exists a fixed ratio between the input and output sequence lengths, CNNs are a viable alternative.

The inductive bias of translational equivariance in CNNs is also typically advantageous for time series tasks. This is because similar sequence components and events are likely to occur at arbitrary positions along the time series and should be detected independently of their location. Additionally, CNNs, like RNNs, can learn long-range temporal features when designed with a sufficiently deep architecture. In the early layers of the network, time series data are processed locally, while the receptive field (the region of the input sequence from which values affect a convolution operation at a given layer) grows polynomially with depth. The precise theoretical growth of the receptive field is determined by factors such as kernel sizes in each convolution layer, the stride of convolution operations, the width of any potential pooling operations, and the possible use of dilated convolutions. Dilated convolutions involve, figuratively, the insertion of spacing between kernel weights to convolve over greater distances in the input. It is important to note that the effective receptive field is always smaller than the theoretical (see Luo et al. 2017). See Paper C of Part III for an example of how these parameters can expand the receptive field of an FCN. Ultimately, convolutions in the deep layers of a CNN can extract complex features (due to the hierarchical feature learning properties of CNNs, as

previously discussed in the context of images) spanning long ranges in the input sequence. In practice, it is often challenging to optimize RNNs to respond to very long-range dependencies (Yoshua Bengio, Simard, et al. 1994). Additionally, CNNs, as feed-forward networks, are generally easier to optimize compared to RNNs, which frequently encounter vanishing and exploding gradient problems (Pascanu et al. 2013). Lastly, CNNs tend to train more quickly due to their non-sequential computations, which permit greater parallelization, unlike the intrinsically sequential nature of RNNs.

In summary, the inductive biases of CNNs, which make them well-suited for image processing, also apply to several time series tasks (Bai, Z. Kolter, et al. 2018; Zhuang Liu et al. 2022). Given their proven track record in image segmentation, we hypothesised that FCNs could also serve as strong candidate models for the time series equivalent of sleep staging. However, a critical technical difference between EEG time series segmentation for sleep staging and medical image segmentation lies in the fact that sleep stages usually span multiple input time points (with one stage typically scored for every 30 seconds of signal, which may be sampled at hundreds of Hertz), while image segmentation masks generally have a direct point-to-point correspondence with the input image. In Part III, Paper C, we introduce and discuss solutions to bridge this gap.

3.3.7 High-frequency sleep staging

As outlined above, both the original R&K- and current AASM visual sleep scoring guidelines are limited to scoring sleep into discrete, contiguous blocks of (typically) 30 seconds of length, thus not fully accounting for the continuous and complex nature of the underlying brain physiology. Most automatic sleep scores have inherited this restriction because they were rule-based systems designed to mimic these standards as closely as possible or because expert-generated labels derived from humans following these same rules were used to train a supervised machine learning classification model.

Numerous sleep and brain disorders are known to impact sleep patterns on short time scales, also called sleep microarchitecture. These disorders include narcolepsy (Ferri, Miano, et al. 2005), Parkinson's disease (Priano et al. 2019), sleep-disordered breathing (Chan et al. 2020; Kheirandish-Gozal et al. 2007), epilepsy (Halász et al. 2002), and others, which affect, for example, the cyclic alternating pattern (CAP) – a key component of sleep microarchitecture characterized by periodic EEG activity in NREM sleep with alternating phases of brain activation (Parrino et al. 1998; Terzano et al. 2002). It is, therefore, reasonable to hypothesize that a sleep staging model capable of extracting sleep stages at higher frequencies could serve as a valuable tool for detecting novel sleep (EEG) biomarkers (Péter Halász et al. 2004; Hasan 1983; H. Koch, Poul Jennum, et al. 2019). In a broader context, there is growing interest in employing machine learning or other data-driven approaches to discover new representations of sleep that do not reduce the intricate physiology of sleep into discrete

blocks, thereby promoting the identification of novel biomarkers. One example is using "hypnodensity" plots, probabilistic-like hypnograms of epoch and stage-wise confidence scores generated by a machine learning model, as suggested by Stephansen et al. (2018) for diagnosing narcolepsy patients.

Some of the earliest works on automatic sleep staging, such as those based on spectral analysis outlined above, were designed to overcome precisely the issue of fixed-length window segmentation. So-called *adaptive segmentation* dynamically shrinks or expands the scoring windows to match relevant parts of the input that adheres to some rules concerning the signal characteristics. For instance, a moving window may be compared to a stationary window positioned at the beginning of a new segment, and once a certain measure of dissimilarity between the two windows crosses a pre-defined threshold, a new segment is defined. These methods were studied in depth and often among the primary motivations behind developing an automated stager in the first place (St. Kubicki et al. 1989). Creutzfeldt et al. (1985), although for the more general analysis of clinical EEG, not specifically during sleep, further went on to detect groups of similar (adaptively detected) segments using unsupervised cluster analysis techniques. In principle, similar techniques could be applied to sleep EEG to detect adaptively sized sleep stage segments and redefine a new set of sleep stages in an unsupervised data-driven manner, for instance, in critically ill patients where the normal five sleep stages no longer apply or are difficult to score. However, both the adaptive segmentation and subsequent cluster analysis depend on several parameters that will ultimately control the number and nature of discovered groups of stages, which may be difficult to determine objectively without a solid physiological hypothesis. This is not often the case in many patient groups where sleep stages are poorly defined and understood. More broadly, the definition of sleep and how it is studied and diagnosed relies on the current sleep scoring standard, which defines five stages in 30-second blocks. Going towards shorter segments and potentially other stages is thus a fundamentally complex problem.

Since the advent of machine learning, most attempts to archive high-frequency sleep staging have included shortening the time window for feature extraction or applying neural network classifiers in sliding windows over the signal. Both methods have limitations. In the former case, the computed features, e.g., time- and frequency domain features, may be noisy or ill-defined at shorter time scales and do not account for potential long-range dependencies. The latter sliding window approach may smooth out stage transition boundaries and transient sleep stages. A better approach would be to train supervised models on manually defined labels scored in shorter segments. However, this approach is also fundamentally complex because humans remain restricted to scoring according to the AASM guidelines.

None of the described approaches has convinced the medical community to break away from fixed-length scoring, and manual scoring, even today, remains limited to 30-second intervals. In Part III, Papers C and D, we propose another way to model high-frequency sleeping patterns by fitting

fully convolutional neural networks that, per default, score in 30-second intervals and are supervised by the standard 30-second labels yet learn an implicit, intermediary representation of sleep, from which scores may be generated at inference time at a higher frequency. We further study these high-frequency stages' potential clinical and diagnostic information in Abstract E and Manuscript F.

3.3.8 Spatial sleeping patterns

Traditionally, sleep has been regarded as a global and top-down brain phenomenon in which a sleep regulatory network imposes sleep on the entire brain. This perspective is reflected in the visual scoring of sleep stages, where a single sleep stage is assigned to each segment of brain and body recordings. The concept of sleep as a global state has been, and remains, essential for understanding sleep in both health and disease. This is because global metrics, such as total sleep time, sleep efficiency, and the percentage of time spent in each sleep stage, tend to exhibit systematic variations that establish well-defined diagnostic criteria for various sleep disorders (Sateia 2014).

Global sleep, however, is likely an oversimplification of the underlying brain and body physiology of sleep. Accumulating evidence suggests that sleep can also be regarded as a local or spatial phenomenon, wherein individual neuronal networks may exhibit distinct activity and sleep patterns, allowing different regions of a sleeping individual's brain to potentially be in different sleep stages simultaneously (J. Krueger et al. 2019; J. M. Krueger et al. 2008; Siclari et al. 2017). From the local sleep perspective, the characteristics of whole-organism level sleep arise from synchronizing the states of multiple local networks.

Local sleep is a well-established phenomenon in other species, such as the characteristic unihemispheric slow-wave sleep observed in dolphins (Mukhametov et al. 1977) and several bird species (Mascetti 2016), where one hemisphere of the brain sleeps while the other remains awake. Similar observations have been made in rats, where local cortical areas enter slow-wave sleep after extended wakefulness (Vyazovskiy et al. 2011). In humans, regional differences in slow-wave activity have been observed in EEG recordings along the anteroposterior cortical axis (De Gennaro, Ferrara, Curcio, et al. 2001; De Gennaro, Ferrara, Vecchio, et al. 2005). Sleep spindles, another marker of non-REM sleep, have also been found to occur locally (Nir et al. 2011). Overall, local sleep may offer valuable insights for understanding sleep disorders. For example, insomnia disorder has been proposed to involve a desynchronization of local sleep-wake activity, with some regionally specific neuronal structures exhibiting wake-like activity while others display NREM activity (Buysse et al. 2011).

Because the U-Sleep automatic sleep staging model, described in Paper D of Part III, was trained to score sleep stages using any EEG electrode as input, it raises the intriguing question of whether spatial sleep staging patterns can be observed by applying the model to various physical EEG electrode positions. It is important to note that the U-Sleep model was not explicitly designed for

this purpose. On the contrary, it was optimized to induce *invariance* to channel EEG positions, outputting the same sleep stage regardless of the EEG input. However, this invariance was only explicitly imposed when scoring in typical 30-second intervals. Manuscript F of Part III demonstrates that the similarity between sleep stages scored by U-Sleep in different EEG electrodes decreases as the frequency of stages increases. Manuscript F and the Future Directions chapter 15 discuss the potential and limitations of studying spatial sleep patterns as output by models like U-Sleep.

Chapter 4

Summaries of papers and manuscripts

4.1 Medical image segmentation

Our work on medical image segmentation is presented in Part II. The primary goal was to develop a machine learning pipeline (a model and its optimization procedure) that can be transferred between medical image segmentation tasks, clinical cohorts and scanners without requiring manual hyperparameter re-configurations. Such a system would be clinically robust according to the first definition of section 2.1 because it can easily be re-trained to any specific clinical task of interest without requiring expert knowledge of how to develop machine models. Part II contains two published manuscripts, summarized below in chronological order.

4.1.1 Summary of Paper A

One Network To Segment Them All: A General, Lightweight System for Accurate 3D Medical Image Segmentation (Perslev, Dam, et al. 2019), was published at the 2019 Medical Image Computing and Computer Assisted Intervention (MICCAI) conference. In this work, we developed the Multi-Planar U-Net (MPUNet), a fully convolutional neural network (FCN) based on the U-Net architecture for cross-task medical image segmentation (Falk et al. 2019; Ronneberger et al. 2015). The defining feature of the MPUNet is its multi-planar data augmentation technique, in which large sets of 2D image planes are sampled from several random view orientations onto the 3D data to train a parameter-efficient 2D model while utilizing most of the available information in the labelled training data. Using the induced rotational equivariance property of the augmentation scheme, the MPUNet can be applied multiple times along different views when segmenting new scans to establish a single-model ensemble-like prediction of higher quality.

The MPUNet was developed to participate in the 2018 Medical Segmentation Decathlon (MSD) challenge (see section 3.2.4 and Simpson et al. 2019) in which participating teams should develop machine learning pipelines that could automatically solve ten highly variable segmentation tasks across MRI and CT scans. The MPUNet, although using a fixed model architecture and hyperparameter-set and thus not relying on compute-intensive automatic hyperparameter experiments, ranked 5^{th} and 6^{th} place in the first- and second phases of the challenge, respectively. Empirically, we found the MPUNet robust across tasks even without hyperparameter modifications because it rarely overfitted due to the multi-planar data augmentation mechanism. Consequently, it was possible to use a single

fixed model with a high approximation capacity, which could solve most medical segmentation tasks, yet rarely encounter significant overfitting issues. This strategy was to utilize that even small labelled datasets of medical image volumes contain large amounts of information compared to the number of free parameters of a 2D model, which operates on input data sampled at a lower dimensionality.

4.1.2 Summary of Paper B

Cross-Cohort Automatic Knee MRI Segmentation with Multi-Planar U-Nets (Perslev, Pai, et al. 2022) was published in the Journal of Magnetic Resonance Imaging. While the MPUNet model was developed and evaluated for cross-task performance in Paper A, Paper B assessed the performance of the MPUNet for the specific task of knee cartilage segmentation but across clinical cohorts and MRI scanner sequences. The MPUNet was applied without further modifications to its architecture or hyperparameters across three clinical cohorts imaged under different MRI protocols. It was compared to the previously evaluated Knee Imaging Quantification (KIQ) framework and a state-of-the-art deep learning-based system for knee cartilage segmentation. It matched or exceeded their performances for all cohorts. The MPUNet was also evaluated for cross-cohort training. A single instance of the model was trained on images from two cohorts of variable MRI sequences simultaneously without losing significant performance on the individual cohorts. The last experiment highlighted a promising path to obtaining clinically robust segmentation systems by training on large and varied datasets pooling multiple cohorts imaged under different protocols, an idea we also pursued in sleep staging in Paper D below.

4.2 Sleep staging

Our work on automatic sleep staging is presented in Part III. The primary goal was to develop a machine learning model that accepts as input variable PSG data from arbitrary clinical cohorts, clinical sites and data recording pipelines and outputs an (optionally high-frequency) hypnogram as accurate as those scored by human experts. Part III contains two published manuscripts, one published conference abstract and one non-published manuscript, summarized below in chronological order.

4.2.1 Summary of Paper C

U-Time: A Fully Convolutional Network for Time Series Segmentation Applied to Sleep Staging (Perslev, Jensen, et al. 2019), was published at the 2019 Conference on Neural Information Processing Systems. In this work, we proposed the *U-Time* model, an FCN for time series segmentation based on the U-Net architecture initially proposed for image segmentation. U-Time maps input sequences of

arbitrary length to sleep stages of variable frequency. This is done by implicitly classifying each input signal from which final output stages are produced by aggregating these high-frequency scores over fixed-length intervals. Based on the proven ability of the U-Net to learn highly variable image segmentation tasks and our findings in Papers A and B (see Part II), we hypothesized that an FCN-based sleep staging model could be a robust candidate model type for clinical cross-cohort sleep staging. FCNs are generally able to learn complex segmentation functions, and being fully feedforward architectures easier to optimize than, for instance, recurrent neural networks, which were typically used at the time. As such, a primary focus of the work was on the so-called hyperparameter stability of the proposed model. That is, to evaluate the performance of a U-Time model with fixed hyperparameters when trained and assessed across several distinct clinical cohorts. Across seven datasets, U-Time reached a high performance comparable to other models that were hyperparameter-tuned explicitly for the individual datasets. U-Time was compared to the state-of-the-art DeepSleepNet model (a mixed RNN+CNN model) when also applied with fixed hyperparameters (Supratak et al. 2017). The CNN+RNN model was more sensitive (more significant drops in performance) to cohort changes and would thus require manual tuning to be used in practice.

In conclusion, Paper C showed that FCN models are suitable for time series segmentation tasks such as sleep staging and may be more hyperparameter-stable than other sequence models that were popular at the time, such as CNN+RNN models. In combination with the ability of U-Time to output scores at a higher frequency at prediction time, although not quantitatively studied in Paper C, U-Time seemed a strong candidate for developing a clinically robust sleep staging model, which could overcome several of the limitations of the preceding automatic sleep stage systems described in the background material in Part 3.3 below.

4.2.2 Summary of Paper D

U-Sleep: Resilient High-Frequency Sleep Staging (Perslev, Darkner, et al. 2021), was published at npj Digital Medicine in 2021. The paper's primary objective was to extend upon the findings of Paper C and train a clinically robust version of the U-Time model, which can be applied to variable PSG data from arbitrary clinical cohorts, clinical sites and data recording pipelines. The resulting model, called *U-Sleep*, was simultaneously trained and evaluated on PSG data from 15,660 unique patients of 16 independent clinical studies. The resulting training dataset was the most extensive and heterogeneous PSG data ever used to train a machine learning model. It included data from both healthy individuals and (mostly sleep-disordered) patients of diverse demographics in dispersed geographical locations (although mainly in the USA) recorded using variable hardware through several decades (1988 – 2018). Because of the ability of U-Time to perform robustly across datasets, it was hypothesized that a larger version of the U-Time model could learn a suitable average representation of sleeping

patterns that may not score perfectly on each dataset but has stable performance in a wide range of practical, clinical scenarios.

The U-Sleep model was trained on randomly sampled batches of data from across all datasets, which were minimally preprocessed, as well as randomly selected input EEG and EOG channel combinations in each batch, without providing information on the origin, channel derivation or other data characteristics to the model, inducing strong invariance towards a wide range of input data variability. Surprisingly, the resulting model could not only score most cohorts accurately using nearly any combination of input EEG and EOG channel derivations but even outperformed multiple other models on two datasets from new clinical sites not seen during training, despite the other models being explicitly trained on data from those sites. When compared to the consensus scores of five human experts, U-Sleep scored at least as accurately as the best of the five experts on healthy and sleep-disordered subjects.

The ability of U-Sleep to score sleep stages at higher frequencies was investigated. Specifically, we tested if scores computed at higher frequencies carried additional information allowing the separation of healthy individuals from patients with obstructive sleep apnea. By computing sleep stage transition triplet frequency features and fitting a random forest model, we found that the classification performance increased significantly with sleep staging frequency. While not proving a direct connection between the high-frequency outputs of U-Sleep and the underlying sleep physiology, these experiments showed that the scores contain some information which may support future biomarker development.

4.2.3 Summary of Abstract E

Automatic detection of abnormal sleeping patterns in stroke patients using high-frequency sleep staging (Perslev, West, et al. 2022), was published in the 2022 Sleep Europe conference proceedings. A set of preliminary experiments were conducted with the U-Sleep model developed in Paper D to investigate its performance on acute stroke patients and if its high-frequency sleep stages contain additional information to separate acute stroke patients from the control of healthy and sleep-disordered individuals. While U-Sleep was extensively evaluated in Paper D, the datasets did not contain many PSGs from patients with severe brain disorders like stroke. Because EEG recordings from acute stroke patients are significantly different from the regular and more complex, scoring sleep stages in stroke patients, even for human experts, is challenging (Cohn et al. 1948; Jordan 2004). Evaluating U-Sleep on stroke patients was thus performed to investigate its worst-case performance.

Across 233 PSGs from stroke patients in the acute- or sub-acute phase, U-Sleep scored stage Wake similarly to healthy individuals, whereas stage REM was scored with below typical performance. Non-REM stage sleep was scored accurately when grouping stages N1, N2, and N3 but significantly

less accurately on the individual stages. This was, however, in some cases due to human annotators' inability to score the separate stages.

In line with the observations of Paper D, high-frequency sleep stages facilitated more easily separation of stroke patients from the control cohort, strengthening the hypothesis that U-Sleep's high-frequency scores contain information lost in the typical 30-second scores. However, as in Paper D, these experiments did not reveal if the scores reflect underlying physiology or if their variations between cohorts represent, for instance, model uncertainty on complex cases.

4.2.4 Summary of Manuscript F

U-Sleep v2: Single-Channel, High-Frequency, and Spatial Sleep Staging for Complex EEG, has not yet been published. The manuscript makes several contributions to address further the clinical robustness and applicability of the U-Sleep model developed in Paper D.

In the first part, new versions of U-Sleep were introduced, called U-Sleep v2, which were trained on an extended dataset with three new, large cohorts. A total of 25,805 PSGs were used for training and evaluating U-Sleep v2. In addition, while U-Sleep v1 and v2 both require (at least) one EEG and (at least) one EOG channel input, two single-channel versions were made that require only (at least) one EEG or (at least) 1 EOG. U-Sleep v2 was as accurate or more accurate than the v1 model on the original datasets also considered in Paper D while scoring the three new cohorts accurately, showing a further generalization ability. In addition, the U-Sleep v2 model produced more accurate estimates of a critical sleep metric, the so-called REM latency, than U-Sleep v1. The single-channel models performed nearly as well as the EEG+EOG counterpart and were as accurate as the best human expert of a group of five, indicating a potential use for wearable devices and in-home sleep studies.

In the second part, the new U-Sleep v2 models were evaluated on five new clinical cohorts of patients with narcolepsy, periodic leg movements (PLM), REM behaviour disorder (RBD), Parkinson's disease (PD) and RBD+PD. U-Sleep scored the former three new cohorts with mean F1 scores between 0.74 to 0.80, nearly matching its performance on other, e.g., healthy cohorts. The more complex PD and RBD+PD cohorts were scored with lower performance but still captured the relevant sleep microstructure compared to expert hypnograms. In addition, the performance of U-Sleep v2 was tested on masked input data (imputed with random noise) to study the model behaviour in the simulated case that some sections of the recorded data are missing (e.g. if the electrode falls off or the recording is paused). U-Sleep outputs were relatively stable even when significant parts of the input signals were replaced with random noise, and correct stages could sometimes (but not always) be predicted even within the area of masked data based on contextual information from pre- and proceeding epochs alone.

In the third and final part, it was investigated if the high-frequency sleep stage outputs lead to different estimations of total time slept in each stage and whether the measures obtained at higher frequencies display vary more or less compared to typical scores when the same subject is studied multiple times on different nights. For stages Wake, N1, N2 and N3, the stage duration metrics output by U-Sleep were more consistent between subsequent studies of the same subjects, even at typical scoring frequency compared to human annotators. The similarity increased as the scoring frequency was increased, indicating that the duration estimates are more stable when computed at higher frequencies, which could improve diagnostic accuracy. For stage REM, while the stability of the stage duration estimates also increased with scoring frequency, it was higher for human expert scores compared to scores from U-Sleep at any scoring frequency.

Finally, a pilot study investigated potential spatial sleeping patterns when applying U-Sleep to different EEG electrode positions. As expected, U-Sleep scores similarly in all channels at typical 1/30 Hz frequency, but at high-frequency outputs, the dissimilarity increases and predictions made in nearby electrodes on the same hemisphere were most similar, while predictions made in electrodes far apart on opposite hemispheres where most dissimilar.

In combination, Manuscript F provides further evidence to support the clinical applicability of the U-Sleep for automatic sleep staging by showing high performance on challenging patient cohorts, stable performance also when data is missing, and the potential for more accurate or stable estimates of key diagnostic sleep parameters such as total stage durations and REM latency. Finally, the paper identifies U-Sleep as a potential tool for studying spatial sleeping patterns, although their physiological relevance must be proven.

Part II

Medical Image Segmentation

Chapter 5

Paper A

One network to segment them all: A general, lightweight system for accurate 3D medical image segmentation

Authors

Mathias Perslev[a], Erik B. Dam[a], Akshay Pai[a, b], and Christian Igel[a].

[a]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

[b]Cerebriu A/S, Copenhagen, Denmark

Published

Medical Image Computing and Computer Assisted Intervention (MICCAI), LNCS 11765, pp. 30-38, Springer, 2019. DOI: https://doi.org/10.1007/978-3-030-32245-8_4

5.1 Abstract

Many recent medical segmentation systems rely on powerful deep learning models to solve highly specific tasks. To maximize performance, it is standard practice to evaluate numerous pipelines with varying model topologies, optimization parameters, pre- & postprocessing steps, and even model cascades. It is often not clear how the resulting pipeline transfers to different tasks.

We propose a simple and thoroughly evaluated deep learning framework for segmentation of arbitrary medical image volumes. The system requires no task-specific information, no human interaction and is based on a fixed model topology and a fixed hyperparameter set, eliminating the process of model selection and its inherent tendency to cause method-level over-fitting. The system is available in open source and does not require deep learning expertise to use. Without task-specific modifications, the system performed better than or similar to highly specialized deep learning methods across 3 separate segmentation tasks. In addition, it ranked 5-th and 6-th in the first and second round of the 2018 Medical Segmentation Decathlon comprising another 10 tasks.

The system relies on *multi-planar* data augmentation which facilitates the application of a single 2D architecture based on the familiar U-Net. Multi-planar training combines the parameter efficiency of a 2D fully convolutional neural network with a systematic train- and test-time augmentation scheme, which allows the 2D model to learn a representation of the 3D image volume that fosters generalization.

5.2 Introduction

More and more systems for medical image segmentation rely on deep learning (DL). However, most publications on this topic report performance improvements for a particular segmentation task and imaging modality and use a specialized processing pipeline adapted through hyperparameter tuning. This makes it difficult to generalize the obtained results and bears the risk that the reported findings are artifacts. In line with the idea behind the 2018 Medical Segmentation Decathlon (MSD)[1] (Simpson et al. 2019), a challenge evaluating the generalisability of machine learning based segmentation algorithms, we argue that new segmentation systems should be evaluated across many different data cohorts and maybe even tasks. This reduces the risk of unintentional method overfitting and may help to gain more general insights about, for example, superior model architectures and learning methods for particular problem classes. This does not only contribute to our basic understanding of the segmentation algorithms, but also to the clinical acceptance and applicability of the systems – even if the generality could come at the cost of not reaching state-of-the-art performance on each individual cohort or task.

A DL segmentation framework that works across a wide range of tasks and in which the individual components and hyperparameters are sufficiently understood allows to automate the task-specific adaptations. This is a prerequisite for being useful for practitioners who are not experts in DL. Big compute clusters offer a way to design systems that provide accurate segmentations for a variety of tasks and do not require tuning by DL experts. If compute resources are not limited, automatic model and hyperparameter selection can be implemented. Given new training data, the systems tests a large variety of segmentation algorithms and, for each algorithm, explores the space of the required hyperparameters. While this approach may produce powerful systems, and was employed to variable extents by top-performing MSD submissions, we argue that it has crucial drawbacks. First, it comes with a risk of automated method overfitting, even if the data is handled carefully. Second, the approach may be prohibitive in clinical practice (and for many scientific institutions) when there is simply no access to sufficient (data regulations compliant) compute resources.

This paper presents an open-source system for medical volume segmentation that addresses all the issues outlined above. It relies on a single neural network of fixed architecture that **1)** showed very good performance across a variety of diverse segmentation tasks, **2)** can be trained efficiently without DL expert knowledge, large amounts of data, and compute clusters, and **3)** does not need large resources when deployed. The system architecture is a 2D U-Net (T. Koch et al. 2019; Ronneberger et al. 2015) variant. The decisive feature of our approach lies in extensive data augmentation, in particular by rotating the input volume before presenting slices to the fully convolutional network.

[1] `http://medicaldecathlon.com`

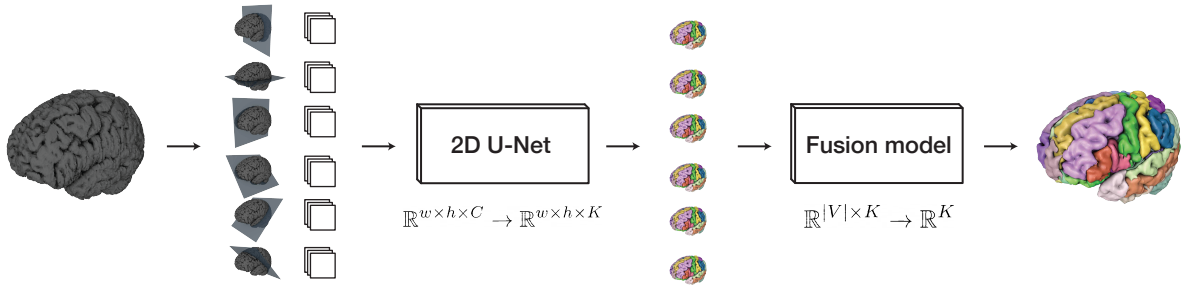


Figure 5.1: Model overview. In the inference phase, the input volume (left) is sampled on 2D isotropic grids along multiple view axes. The model predicts a full volume along each axis and maps the predictions into the original image space. A fusion model combines the 6 proposed segmentation volumes into a single final segmentation.

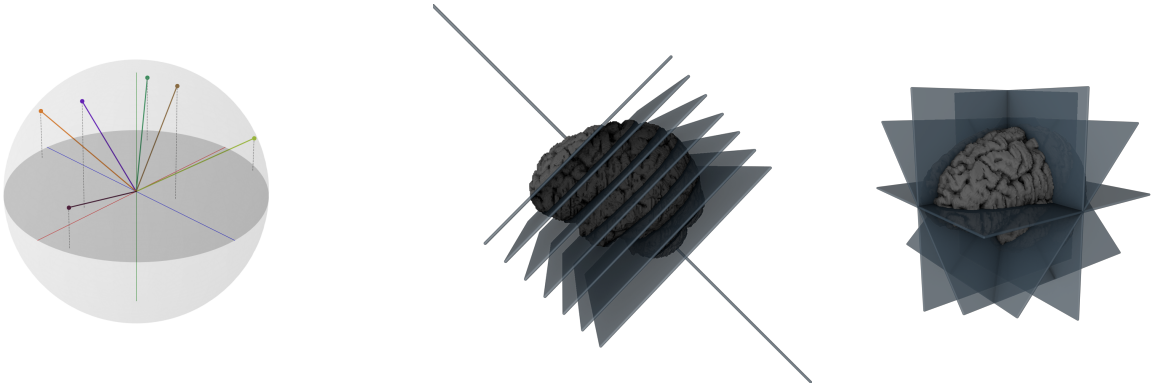
Because of the latter, we refer to our approach as *multi-planar* U-Net training (*MPUnet*). We present a thorough evaluation of our system on a total of 13 different 3D segmentation tasks, including 10 from MSD, on which it obtains high accuracies – often reaching state-of-the-art performance from even highly specialized DL-based methods.

5.3 Method

At the heart of our system lies a 2D U-net (Ronneberger et al. 2015) modified slightly to **1)** include batch normalization layers (Ioffe et al. 2015) intervening each double convolution- and up-convolution block and **2)** use nearest-neighbor up-sampling followed by convolution to implement up-convolutions (Odena et al. 2016). Basic network topology and hyperparameters can bet set to their default choices as done in all experiments in this paper, see Table A.1 in the supplementary material for an overview. Compared to (Ronneberger et al. 2015), the number of filters has been increased by a factor of $\sqrt{2}$, see supplementary Table A.6 for details. As a result, the model has ≈ 62 million parameters. While one would assume that the size of the model is a crucial hyperparameter, we kept the model architecture the same for all tasks. For each task, only the filters in the first layer were resized according to the number C of input channels and the number of output units was set to the number of classes K.

The decisive feature of our multi-planar U-Net training (*MPUnet*) is the generation of the inputs at training and test time, which is done by sampling from multiple planes of random orientation spanning the image volume. That is, the network must learn to segment the input seen from different views, see Fig. 5.1.

The model $f(x;\theta)$ takes as input multi-channel 2D image slices of size $w \times h$, $x \in \mathbb{R}^{w\times h\times C}$, and outputs a probabilistic segmentation map $P \in \mathbb{R}^{w\times h\times K}$ for K classes. Prior to training we define a set $V = \{v_1, v_2, ..., v_i\}$ of i randomly sampled unit vectors in $\mathbb{R}^3$. The set defines the axes through the image volume along which we sample 2D inputs to the model, visualized in Fig. 5.2.



(a) (b) (c)

Figure 5.2: (a) Visualization of a set V of sampled view axis unit vectors. (b) Illustration of images sampled along one view. (c) Illustration of multiple images sampled along multiple unique views.

We re-sample the set V until all pairs of vectors have an angle of at least 60 deg between them. A sampled set of planar axes is shown in Fig. 5.2a. Note that the model could also be fit using a set of fixed, predefined planes, but we found no performance gain in doing so, even if the fixed set included the standard planes. We use $i = 6$ for all reported evaluations. This number was chosen based on prior experiments in which we observed monotonically improving performance with the inclusion of additional planes and $i = 6$ providing a good balance between accuracy and computation, see supplementary Table A.2.

During training, the model is provided batches of images randomly sampled from the i planes in V without supplying information about the corresponding axis. During inference, the model predicts along each plane producing a set of i segmentation volumes $\mathbf{P} = \{P_v \in \mathbb{R}^{w \times h \times d \times K} \mid v \in V\}$. Each P_v is mapped to the input image space to obtain point correspondence by assigning to each voxel in the input image the value of its nearest predicted point in P_v. Distances are computed in physical coordinates.

At test-time, the learned invariance to orientation is exploited by segmenting the entire volume from each view. This results in several candidate segmentations for each subject, which are combined by a linear fusion model, see Fig. 5.1. We map $\mathbf{P}$ to a single probabilistic segmentation by a weighted sum of the per-class and per-view softmax-scores. For all $w \cdot h \cdot d$ voxels x in $\mathbf{P}$ and each class $k \in \{1, ..., K\}$, the *fusion model* $f_{\text{fusion}} : \mathbb{R}^{|V| \times K} \to \mathbb{R}^K$ calculates $z(x)_k = \sum_{n=1}^{|V|} W_{n,k} \cdot p_{n,x,k} + \beta_k$. Here $p_{n,x,k}$ denotes the probability of class k at voxel x as predicted by segmentation P_n. The $W \in \mathbb{R}^{|V| \times K}$ weighs the probabilities of each class as predicted from each view and $\beta \in \mathbb{R}^K$ are bias parameters, which can adjust the overall tendency to predict a given class. The parameters of f_{fusion} are learned from the validation data. The model scales the predictions according to which views do well on each class, motivated by the fact that different target classes may appear in different shapes and levels of recognizability when seen from the different directions in V.

Isotropic Image Sampling. Interpolation is needed to sample image planes not aligned with the original voxel grid. We use tri-linear and nearest-neighbour interpolation to sample the image and label map, respectively. We take advantage of the necessity for interpolation by sampling images on isotropic grids in the physical scanner space, oriented according to the patient's position in the scanner. This ensures that the model always operates on images in which the shapes of anatomical structures are maintained across scanners and acquisition protocols. Note that this approach may lead to over- or under-sampling along some axes, which may lead to loss of image information or interpolation artefacts. Empirically, however, we found that the benefit of maintaining isotropy outweighed potential drawbacks of interpolation.

We must define a set of parameters restricting the sampling. Specifically, we are free to choose **1)** the pixel dimensions, $q \in \mathbb{Z}^+$ (the number of pixels to sample for each image), **2)** the real-space extent of the image (in mm), $m \in \mathbb{R}^+$, and **3)** the real-space distance between consecutive voxels, $r \in \mathbb{R}^+$. Note that two of these parameters define the third. We restrict our sampling to equal q, m and r for both image dimensions producing squared images. We sample images within a sphere of diameter m centered at the origin of the scanner coordinate system. We employ a simple heuristic that attempts to pick q, m and r so that **1)** the training is computable on our GPUs with batch sizes of at least 8, **2)** r approximately matches the resolution of the images along their highest resolution axis and **3)** the sampled images span the entirety of the relevant volume of all images in the dataset. When this is not possible, the requirements are prioritized in the given order, with 1 having highest priority. Note that 3 becomes less important with increasing numbers of planes as voxels missed in one plane are likely to be included in some of the others.

Augmentation. Processing the input image from different views has the the same effect as applying affine transformations to the 3D input and presenting the transformed images to a (single-view) network. Thus, at the heart the MPUnet is a U-Net with extensive, systematic affine data augmentation. On top of the multi-view sampling, we also employ non-linear transformations to further augment the training data. We apply the Random Elastic Deformations algorithm (Simard et al. 2003) to each sampled image in a batch with a probability of 1/3. The elasticity constants σ and deformation intensity multipliers α are sampled uniformly from $[20, 30]$ and $[100, 500]$, respectively. This generates augmented images with high variability in terms of both deformation strength and smoothness.

The augmented images do not always display anatomically plausible structures. Yet, they often significantly improve the generalization especially when training on small datasets or tasks involving pathologies of highly variable shape. However, we weigh the loss-contribution from augmented images by 1/3 in order to optimize primarily over true images.

Pre- and post-processing. Our model uses a minimum of image processing outside of the network itself. We restrain from applying any post-processing of the model's output, because post-processing is typically highly task-specific. We only apply an image- and channel-wise outlier-robust pre-possessing that scales intensity values according to the median and inter-quartile range computed over all non-background voxels. Background voxels are defined by having intensities less than or equal to the first percentile of the intensity distribution.

Implementation. The MPUnet is available as open-source. The fully autonomous implementation makes the MPUnet applicable also for users with limited deep learning expertise and/or compute resources. A command line interface supports fixed split or cross-validation training and evaluation on arbitrary images. Any non-constant hyperparameter can automatically be inferred from the training data. See the GitHub repository at `https://github.com/perslev/MultiPlanarUNet` for a user guide.

5.4 Experiments and Results

We applied the MPUNet without task-specific modifications to a total of 13 segmentation tasks. Ten of those datasets were part of the 2018 MSD challenge, described in detail and sourced on the challenge's website. The remaining three datasets were the MICCAI 2012 Multi-Atlas Challenge (MICCAI) dataset (D. S. Marcus et al. 2007), the EADC-ADNI Harmonized Hippocampal Protocol (HarP) dataset (Boccardi et al. 2015) and a knee MRI dataset from the Osteoarthritis Initiative (OAI) (Dam, Lillholm, et al. 2015). The evaluation covers healthy and pathological anatomical structures, mono- and multi-modal MR and CT, and various acquisition protocols. The mean per-class F1 (dice) scores of the MPUNet are reported in Table 5.1. Note that in MSD tumour segmentation tasks 3 & 7 both organ and tumour are segmented, and the mean F1 for those tasks is lifted by the performance on the organ and decreased by the performance on the tumour. We refer to the supplementary Table A.4 for detailed per-class scores for the ten MSD tasks.

The MPUnet reached state-of-the-art performance for DL methods on the three non-challenge datasets (MICCAI, HaRP and OAI) despite comparable methods being developed and tuned specifically to the cohorts and tasks. On MICCAI, with a mean F1 of 0.74 the MPUnet compares similar to the 0.74 obtained in (Moeskops et al. 2016) using a 2D multi-scale CNN on brain-extracted images and 0.75 obtained in (Ganaye et al. 2018) using a combination of a multi-scale 2D CNN, 3D patch-based CNN, a spatial information encoder network and a probabilistic atlas also on brain-extracted images. With a mean F1 of 0.85 on HarP, the MPUnet compares favorable to 0.78-0.83 (depending on subject disease state) reported in (Roy et al. 2018). On OAI, with a mean F1 of 0.87, the MPUnet

Table 5.1: Performance of the MPUnet across thirteen segmentation tasks. The shown F1 (dice) scores are mean values computed across all non-background per-class F1 scores. For the 10 MSD datasets evaluation was performed by the challenge organisers on non-publicly available test-sets. For MICCAI and HarP, evaluation was performed over three trials. Five fold cross-validation was used for OAI. The 'Classes' column include the background class, which is not included when computing the F1 scores. The 'Size' column gives the total dataset size. Note that the F1 standard deviations for tasks 8, 9 & 10 are not yet published by the challenge organizers. We refer to `http://medicaldecathlon.com/results.html` for a detailed comparison of our results (team CerebriuDIKU) with those of other challenge participants.

	Dataset	Modality	Segmentation Target(s)	Classes	Size	F1 Score
	MICCAI	MRI	Whole-Brain	135	35	0.74 ± 0.03
	HarP	MRI	L+R Hippocampus	3	135	0.85 ± 0.03
	OAI	MRI	Knee Cartilages	7	176	0.87 ± 0.06
2018 Medical Segmentation Decathlon	Task 1	MRI	Brain Tumours	4	750	0.60 ± 0.24
	Task 2	MRI	Cardiac, Left Atrium	2	30	0.89 ± 0.09
	Task 3	CT	Liver & Tumour	2	201	0.76 ± 0.18
	Task 4	MRI	Hippocampus ROI.	2	394	0.89 ± 0.04
	Task 5	MRI	Prostate	3	48	0.78 ± 0.10
	Task 6	CT	Lung Tumours	2	96	0.59 ± 0.23
	Task 7	CT	Pancreas & Tumour	3	420	0.48 ± 0.21
	Task 8	CT	Hepatic Ves. & Tumour	3	443	0.49
	Task 9	CT	Spleen	2	61	0.95
	Task 10	CT	Colon Cancer	2	190	0.28

gets near the 0.88/0.89 (baseline/follow-up) obtained in (Ambellan et al. 2019) using a task-specific pipeline including 2D- and 3D U-nets along with multiple statistical shape model refinement steps. However, the comparison cannot be directly made as (Ambellan et al. 2019) worked on a smaller subset of the OAI data and predicted only 4 classes while we distinguished 7.

The MPUnet ranked 5th and 6th place in the first and second phases of the Medical Segmentation Decathlon respectively, in most cases comparing unfavorable only to significantly more compute intensive systems (see below).[2]

The question arises how the performance of a 2D U-net with multi-planar augmentation compares to a U-net with 3D convolutions. Such 3D models are computationally demanding and typically need – in our experience – large training datasets to achieve proper generalization. While we are not making the claim that the MPUnet is universally superior to 3D models, we did find the MPUnet to outperform a 3D U-net of comparable topology, learning and augmentation procedure across multiple tasks including one for which the 3D model had sufficient spatial extent to operate on the entire input volume at once. We refer to the supplementary Table A.5 for details. We also found the MPUnet superior to both single 2D U-Nets trained on individual planes as well as ensembles of separate 2D U-Nets trained on different planes, see Table A.2 & A.3 and Fig. A.1 in the supplementary material.

[2]For comparison, the median F1 scores over all 10 tasks of the best five phase 1 submissions were 0.74, 0.67, 0.69, 0.66, and (our method) 0.69. Note that the official ranking was based on a more rigorous statistical analysis.

5.5 Discussion and Conclusions

The empirical evaluation over 13 segmentation tasks showed that multi-planar augmentation provides a simple mechanism for obtaining accurate segmentation models without hyperparameter tuning. With no task-specific modifications the MPUnet performs well across many non-pathological tissues imaged with various MR and CT protocols, in spite of the target compartments varying drastically in number, physical size, shape- and spatial distributions, as well as contrast to the surrounding tissues. Also the accuracies on the more difficult pathological targets are favorable compared to most other MSD contesters.

The MSD winning algorithm (Isensee et al. 2018) relied on selecting a suitable model topology and/or cascade from an ensemble of candidates through cross-validation. In contrast to this and other top-ranking participants, we were interested to develop a task-agnostic segmentation system based on a single architecture and learning procedure that makes the system lightweight and easily transferable to clinical settings with limited compute resources.

That the MPUnet can be applied 'as is' across many tasks with high performance and its robustness against overfitting can be attributed to both the fully convolutional network approach, which is already known to generalize well, and our multi-planar augmentation framework. The latter allows us to apply a single 2D model with fixed hyperparameters, resulting in a fully autonomous segmentation system of low computational complexity. Multi-planar training improves the generalization performance in several ways: **1)** Sampling from multiple planes allows for a huge number of anatomically relevant images augmenting the training data; **2)** Exposing a 2D model to multiple planes takes the 3D nature of the input into account while maintaining the statistical and computational efficiency of 2D kernels; **3)** The systematic augmentation scheme allows test time augmentation to be performed, which increases the performance through variance reduction if errors across views are uncorrelated for a given subject (visualized in supplementary Fig. A.2). This makes the MPUnet an open source alternative to 3D fully convolutional neural networks.

Acknowledgements

We would like to thank both Microsoft and NVIDIA for providing computational resources on the Azure platform for this project.

Chapter 6

Paper B

Cross-cohort automatic knee MRI segmentation with multi-planar U-nets

Authors

Mathias Perslev[a], Akshay Pai[a, b], Jon Runhaar[c], Christian Igel[a] and Erik B. Dam[a].

[a]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

[b]Cerebriu A/S, Copenhagen, Denmark

[c]Erasmus MC, Rotterdam University, Rotterdam, Netherlands

Published

Journal of Magnetic Resonance Imaging, 55(6):1650-1663, 2022. DOI: `https://doi.org/10.1002/jmri.27978`

6.1 Abstract

Background: Segmentation of medical image volumes is a time-consuming manual task. Automatic tools are often tailored towards specific patient cohorts, and it is unclear how they behave in other clinical settings.

Purpose: To evaluate the performance of the open-source Multi-Planar U-Net (MPUnet), the validated Knee Imaging Quantification (KIQ) framework, and a state-of-the-art 2D U-Net architecture on three clinical cohorts without extensive adaptation of the algorithms.

Study Type: Retrospective cohort study.

Subjects: 253 subjects (146 females, 107 males, ages 57±12 years) from three knee osteoarthritis (OA) studies (CCBR, OAI & PROOF) with varying demographics and OA severity (64/37/24/53/2 scans of KL-grades 0-4).

Field Strength / Sequence: 0.18T, 1.0T/1.5T & 3T sagittal 3D fast spin echo T1w and DESS sequences.

Assessment: All models were fit without tuning to knee MRIs with manual segmentations from three clinical cohorts. All models were evaluated across KL-grades.

Statistical Tests: Segmentation performance differences as measured by Dice coefficients were tested with paired, two-sided Wilcoxon signed-rank statistics with significance threshold $\alpha = 0.05$.

Results: The MPUnet performed superior or equal to KIQ and 2D U-Net on all compartments across three cohorts. Mean Dice overlap was significantly higher for MPUnet compared to KIQ and U-Net on CCBR (0.83 ± 0.04 vs 0.81 ± 0.06 and 0.82 ± 0.05), significantly higher than KIQ and U-Net OAI (0.86 ± 0.03 vs 0.84 ± 0.04 and 0.85 ± 0.03), and not significantly different from KIQ while significantly higher than 2D U-Net on PROOF (0.78 ± 0.07 vs 0.77 ± 0.07, $P = 0.10$, and 0.73 ± 0.07). The MPUnet performed significantly better on $N = 22$ KL-grade 3 CCBR scans with 0.78 ± 0.06 vs 0.75 ± 0.08 for KIQ and 0.76 ± 0.06 for 2D U-Net.

Data Conclusion: The MPUnet matched or exceeded the performance of state-of-the-art knee MRI segmentation models across cohorts of variable sequences and patient demographics. The MPUnet required no manual tuning making it both accurate and easy-to-use.

6.2 Introduction

Recent advances in machine learning have pushed automatic segmentation tools close to human performance for medical image analysis (Litjens et al. 2017; Shen et al. 2017). This includes the automatic quantification of cartilage compartments from magnetic resonance imaging (MRI) scans, which facilitates robust, large-scale quantitative studies of osteoarthritis (OA) (Gan et al. 2020). Until recently, most validated automatic segmentation software, such as the Knee Imaging Quantification (KIQ) framework, were specialized and relied at least partially on task-specific knowledge (Dam, Runhaar, et al. 2018). Gan et al. (2020) review a range of successful classical approaches based on, e.g., random forests, deformable models, graph-based algorithms and atlas registration (Ababneh et al. 2011; Dam, Lillholm, et al. 2015; Kashyap et al. 2018; Seim et al. 2010). With advances in deep learning it is now possible to create automatic segmentation models given sufficient training examples alone (LeCun, Bengio, et al. 2015). Numerous deep learning based approaches have been suggested in recent years alone. The majority consider models from the family of fully convolutional networks (FCN) in the popular encoder-decoder architecture, typically inspired by the U-net (Long et al. 2014; Norman et al. 2018; Panfilov et al. 2019; Wirth et al. 2021). The FCN-centered methods for knee MRI segmentation vary in complexity, ranging from using a single-stage U-net to combining several U-nets (e.g., both 2D and 3D) and shape model refinement steps (e.g., a 2D U-net followed by shape model refinement used to identify regions of interest which are then segmented by a 3D U-net followed by another shape model refinement step) (Ambellan et al. 2019; Norman et al. 2018; Panfilov et al. 2019; Tack et al. 2018; Zhou et al. 2018). Different strategies have been employed to render deep learning on 3D data efficient and to cope with limited training data. To increase the efficiency of 3D FCNs, it has been suggested to operate on overlapping patches or on down-sampled scans (Raj et al. 2018). Another strategy is to employ a 2D FCN to segment each scan slice independently (Norman et al. 2018; Panfilov et al. 2019; Wirth et al. 2021; Zhou et al. 2018). The 2D approach has been extended in different ways including considering multiple planes or 3D surface model optimization schemes (Ambellan et al. 2019; Hyungjin Lee et al. 2018; F. Liu et al. 2018; Prasoon et al. 2013; Tack et al. 2018; Zhou et al. 2018).

Despite a vast number of existing deep learning based methods for OA segmentation (often shown to perform accurately compared to human annotators), no method has seen widespread clinical adaptation. While such adaptation is complex due to both practical, ethical, and legal factors, central research problems related to the models themselves also remain. For instance, it is largely unclear how different models and methods compare even on a single cohort. Which type of model should be perused for clinical validation for a given task? Secondly, it is even less clear if one model designed to work well on a single dataset can also be expected to work well in other clinical scenarios,

e.g., on data from new patient cohorts, scanner sequences or scanner manufactures. The 2019 OA MRI segmentation challenge made one attempt towards addressing the former problem (Arjun Desai et al. 2021). A range of deep-learning-based methods were compared and evaluated for knee MRI segmentation on a single cohort. Multiple methods were found to perform at clinically applicable levels. Surprisingly, the challenge demonstrated that even the simple 2D U-Net baseline model was highly competitive (Ronneberger et al. 2015). This result indicates that many deep learning based approaches are viable when tuned to a specific dataset, and that the method need not be very complex. As for the second problem, however, it is unclear how the 2D U-net and other challenge methods (often using more complex setups with, e.g., cascaded models or multiple post-processing steps) would perform when trained on new clinical cohorts (e.g., a smaller set of annotated scans or knees with different levels of OA severity) without re-tuning of the hyperparameters, or when trained across multiple cohorts at once. The robustness of a model under such cross-cohort scenarios is crucial when adapting it for clinical practice, as tuning a neural network model for new data typically requires both compute resources and access to technical experts.

The purpose of this study was to investigate the cross-cohort performance and robustness of state-of-the-art (classical as well as deep learning based) automatic knee segmentation methods. Its primary focus was on the recently proposed Multi-Planar U-Net (MPUnet) model. The MPUnet extends the popular 2D U-Net with a unique data-resampling technique, and has been found able to output accurate segmentations across clinical cohorts (and different segmentation tasks) without hyperparameter-tuning (Perslev, Dam, et al. 2019; Simpson et al. 2019). It scored a top-position in the 2019 OA MRI segmentation challenge and a top-5 position in the 2018 Medical Segmentation Decathlon (Arjun Desai et al. 2021; Simpson et al. 2019). The MPUnet is hyperparameter search free in the sense that the default settings have proven to give good results on variable medical image segmentation tasks, so no machine learning expertise is required to train the MPUnet on new data. These findings indicated that the MPUnet could serve as an accurate, yet easy-to-use tool for robust cross-cohort knee MRI segmentation also in clinics with limited access to technical experts. To test this hypothesis, this study investigated the performance and robustness of the MPUnet as compared to other state-of-the-art models for OA segmentation (classical as well as deep learning based) when applied across cohorts without manual adaptation of model- or optimization hyperparameters. A total of four OA segmentation models were considered:

1. The default MPUnet (Perslev, Dam, et al. 2019). The MPUnet relies on a single 2D U-Net (fully convolutional neural network) model fit to 2D image slices sampled (isotropically) along $V = 6$ viewing planes through the image volume. The amount of training data increases V times, but the different views of a volume are not independent of each other. In this way the

extension of the training data resembles data augmentation (Ioffe et al. 2015). Random elastic deformations are applied to a subset of the sampled images to further augment the training dataset, see Supplementary Figure B.1 (Simard et al. 2003). During optimization, images from all planes are fed to the (a priori plane-agnostic) model without additional information about the corresponding image plane, see Figure 5.2 from Paper I (Chapter 5, published Perslev, Dam, et al. (2019). This training setup forces the model to learn to segment the medical target of interest as seen from multiple views. When segmenting a new scan, the model first predicts along each plane in the isotropic scanner space creating a set of V full segmentation volumes for each input scan. The V segmentation suggestions are combined into one final output using a learned fusion model. The single neural network model thus plays the role of V experts in an ensemble-like method. The approach is illustrated in Figure 5.1 of Paper I (Chapter 5, published Perslev, Dam, et al. (2019). The output of the MPUnet is considered the final segmentation with no post-processing steps applied. The MPUnet and its optimization is described in further detail in the Supplementary Information. Additional technical details are given in Paper I (Chapter 5, published Perslev, Dam, et al. (2019).

2. The MPUnet using only a single view (the sagittal view). This corresponds to training a simple 2D U-Net but using the augmentation strategy (i.e., random elastic deformations) and training pipeline of the MPUnet. This ablatation study tests the effectiveness of including additional views.

3. The validated Knee Imaging Quantification (KIQ) automatic segmentation method (Dam, Lillholm, et al. 2015; Dam, Runhaar, et al. 2018). The KIQ method was developed and extensively validated over many years and is partly based on task-specific knowledge on cartilage segmentation. The framework first aligns the considered scan to a reference knee MRI model using rigid multi-atlas registration. Gaussian derivative features are then computed within regions-of-interest for each segmentation compartment individually. The computed features support voxel-wise classifications using compartment specific classifiers, and largest connected component analysis is used to select final segmentation volumes for each compartment.

4. A 2D U-Net as implemented by Panfilov et al., 2019 which represents state-of-the-art performance on the Osteoarthritis Initiative (OAI) dataset (Panfilov et al. 2019), see Methods. The optimization hyperparameters, including loss function, learning rate, weight decay, batch size, number of epochs, etc., have been tuned for the OAI dataset, and this comparison thus allows to study how the popular 2D U-Net transfers to other datasets without re-tuning of its hyperparameters. The Panfilov 2D U-Net performs slice-wise segmentation in the sagittal view without post-processing of the obtained masks. Random augmentations, such as gamma

corrections, scaling and bilateral filtering, are applied during training.

This study aimed to investigate how each of these models perform when applied with default hyperparameters across three distinct OA cohorts to measure their robustness to scanner- and patient demographic variations. Average model performance across all scans in a cohort and as a function of Kellgren and Lawrence (KL) grades were compared (Kellgren et al. 1957). Finally, the ability of the MPUnet to learn segmentations across multiple cohorts at once was investigated.

6.3 Methods

6.3.1 Cohorts

The performance of all segmentation models were evaluated on three distinct cohorts of MR knee scans:

1. **Osteoarthritis Initiative** (OAI) cohort subset consisting of 88 baseline scans and 88 follow-up scans with approximately 1-year interval. Scans were acquired using a Siemens 3T Trio (Erlangen, Germany) scanner and a sagittal 3D Dual-Echo Steady State (DESS) with water excitation sequence. The cohort consists of 45 males and 43 females of ages 61 ± 10 years and BMIs 31.1 ± 4.6. All enrolled participants either had or were at increased risk of developing OA. OA severity was assessed for 44 baseline scans with 0/2/10/30/2 scans of KL-grades 0-4.

2. **Center for Clinical and Basic Research** (CCBR) consisting of 140 scans from 140 subjects (Dam, Folkesson, et al. 2007). Scans were acquired using a 0.18T Esaote C-Span scanner (Genova, Italy) and a Turbo 3D T1w sequence. The cohort consists of 78 females and 62 males of ages 55 ± 15 and BMIs 25.8 ± 4.0. Enrolled participants had both healthy knees and varying degrees of OA with 50/24/13/22/0 scans of KL-grades 0-4.

3. **Prevention of OA in Overweight Females** (PROOF) consisting of 25 knees imaged with 1.5T Simens Symphony (Erlangen, Germany), 1.5T Siemens Magnetom Essenza (Erlangen, Germany) and 1.0T Phillips Intera (Eindhoven, Netherlands) scanners using a 3D sagittal DESS sequence with water excitation (Runhaar et al. 2015). Women aged 50-60 years with BMI $\geq$ 27 and free of knee OA (according to clinical American College of Rheumatology [ACR] criteria) were included in the original study. The sub-cohort considered here consists of 25 females of ages 56 ± 3 and BMIs 32.2 ± 4.1 and 12/11/1/1/0 scans of KL-grades 0-4.

Cohort statistics are summarized in Table 6.1. MRI sequence details are given in Table 6.2. All MRIs of right knees were mirrored to resemble left knees. Informed consent was given by all

Table 6.1: Overview of study populations. Statistics were computed over 88, 140 and 25 subjects for the OAI, CCBR and PROOF cohorts, respectively. OAI = Osteoarthritis Initiative; CCBR = Center for Clinical and Basic Research; PROOF = Prevention of OA in Overweight Females. Age and BMI shows mean and standard deviation. [a]For OAI, the Tibia bone was only annotated in the 88 baseline scans. In the baseline scans, the Medial & Lateral Femoral Cartilages were annotated separately, whereas in the 88 follow-up scans the Femoral Cartilage was annotated as a single compartment.

Cohort	Scans	Subjects	Compartments	Age (years)	BMI (kg/m^2)	Sex (M/F %)
OAI	176	88	6/8[a]	61 ± 10	31.1 ± 4.6	51/49
CCBR	140	140	2	55 ± 15	25.8 ± 4.0	44/56
PROOF	25	25	6	56 ± 3	32.2 ± 4.1	0/100

participants for inclusion into any of the original study cohorts. All data considered in this study was handled and processed in accordance with the relevant data sharing agreements for each study.

6.3.2 Radiological Assessment and Segmentation

OAI: The tibial medial- and lateral cartilages (TMC & TLC), femoral medial- and lateral cartilages (FMC, FLC), medial- and lateral menisci (MM & LM) and patellar cartilage (PC) were manually segmented in all 176 scans by iMorphics (Manchester, UK). The tibia bone (TB) was further annotated in the 88 baseline scans. In the baseline scans FMC and FLC were annotated separately, whereas in the 88 follow-up scans the femoral cartilage was annotated as a single compartment. KL-grades were assessed for all scans by trained radiologists from The David Felson Lab, School of Medicine, Boston University. CCBR: TMC and FMC were manually segmented in all scans by trained radiologist PCP (Denmark). KL-grades were assessed by PCP for 109 out of the total 140 scans. PROOF: TMC, TLC, FMC, FLC, PC and TB were segmented in all scans by trained radiologist DS of Erasmus Medical Center, Rotterdam Universit. KL-grades were assessed by DS for all scans.

Segmentation results on TB are reported in Results but not further discussed, because the compartment is easily segmented by all considered methods.

6.3.3 Segmentation Models

Four segmentation models were evaluated on each of the three MRI cohorts:

1. A default MPUnet model using $V = 6$ planar views (see Supplementary Information for details) (Perslev, Dam, et al. 2019)

2. A $V = 1$ MPUnet using only the sagittal view to test the effect of using multiple views

3. The Knee Imaging Quantification (KIQ) automatic segmentation framework (Dam, Lillholm, et al. 2015; Dam, Runhaar, et al. 2018)

Table 6.2: Overview of cohort MRI sequences. [a]Variable slice thicknesses in 0.703-0.938 mm, typically 0.781 mm. [b]Minor variations in Echo/Repetition Times in 11.1-11.4 ms / 22.2-22.6 ms.

Cohort	OAI	CCBR	PROOF
Scanner	Siemens Trio	Esaote C-Span	Siemens Symphony Siemens Magnetom Essenza Phillips Intera
Vendor Location	Erlangen, Germany	Genoa, Italy	Erlangen, Germany Erlangen, Germany Eindhoven, Netherlands
Scan	3D DESS	Turbo 3D T1w	3D DESS
Field Strength (T)	3.0	0.18	1.5 1.5 1.0
Acquisition Time (min)	10	10	5-10
Plane	Sagittal	Sagittal	Sagittal
Fat Suppression	Water Excitation	None	Water Excitation
Field of View (mm)	140	180	160
Number of Slices	160	110	50-62
Voxel Size (mm^3)	$0.700 \times 0.365 \times 0.365$	$0.781^{a} \times 0.703 \times 0.703$	$1.500 \times 0.420 \times 0.420$ $1.500 \times 0.500 \times 0.500$ $1.500 \times 0.625 \times 0.625$ $1.500 \times 0.310 \times 0.310$
Flip Angle (°)	25	40	25
Bit Depth	12	8	12
Echo/Repetition (ms/ms)	4.7 / 16.3	16 / 50	6.0 / 19.5 8.0 / 21.4 11.3 / 22.3[b]

4. A 2D U-Net as implemented by Panfilov et al., 2019 (Panfilov et al. 2019) marking the state-of-the-art in deep learning for knee MRI segmentation

The MPUnet and KIQ framework were applied with default settings across all cohorts. The 2D U-Net was applied with optimization hyperparameters as in Panfilov et al., 2019 (Panfilov et al. 2019) using the codebase (`https://github.com/MIPT-Oulu/RobustCartilageSegmentation`) provided by the authors with the following exceptions: **(1)** the input image sizes were modified from the default 300×300 on the OAI dataset to 256×256 on CCBR (to match the size of those scans) and 336×336 on PROOF (bilinear resampling was used to down-sample PROOF images from their original variable sizes of 320×320, 384×384 or 512×512 depending on scan; the resampled pixel size was set to 0.47×0.47 mm^2). The size of the images input to the 2D U-Net matched those of the MPUnet on corresponding datasets. **(2)** A common batch-size of 32 was used across the datasets (down from 64) to allow the larger 336×336 PROOF images to fit in our GPU memory. **(3)** The learning rate was reduced to 0.0005 (down from 0.001 on OAI) and the number of training epochs increased to 150 (down from 50 on OAI) when training on the PROOF dataset due to severe

overfitting observed using the default parameters on this small dataset. **(4)** Random horizontal flips were disabled in the augmentation pipeline (leaving random gamma corrections, scaling and bilateral filtering) as all MRI images considered here were mirrored to resemble right knees as described above.

6.3.4 Experiments and Statistical Analysis

All models were trained and evaluated on each of the three study cohorts individually. The MPUnet was further evaluated in a cross-cohort setup. The trained models were applied to a subset of the data held out during training to test their generalization properties.

6.3.5 Single Cohort Setup

On CCBR and OAI, all models were trained and evaluated on a fixed dataset split. On PROOF, all models were trained and evaluated in a leave-one-out (LOO) cross-validation (25-fold CV) setup. We considered fixed training/testing splits and cross-validation strategies for each dataset as Dam, Runhaar, et al. (2018). The CCBR dataset was split into 30 training- and 110 evaluation images and the OAI dataset was split into 44 training- and 44 evaluation images. On PROOF, the model was evaluated in a LOO experiment (training 25 model instances each evaluated on a single, held-out testing scan). The MPUnet was further evaluated using larger training datasets facilitated by either a cross-validation setup with more folds (for CCBR and OAI) or through training on additionally images taken from a different dataset (for PROOF). Specifically, we included 88 images taken from the OAI dataset and added them to the training dataset of PROOF to investigate if the publicly available OAI dataset could reduce the need for new manual segmentations when applying the MPUnet on a new cohort.

6.3.6 Cross-Cohort Setup

A single instance of the model was trained on MRIs from the OAI, CCBR and PROOF datasets simultaneously. For OAI and CCBR, we used the same fixed-splits defined above in the single-cohort setup. We also included all 25 PROOF images into the training set to expose the model to as many and variable images as possible. The model was evaluated on the test-set images of CCBR and OAI. We did not evaluate on PROOF images as no fixed dataset split is available for this small dataset. The cross-cohort model segments only the tibial- and femoral medial cartilages, as those are the only two annotated compartments of the CCBR cohort.

6.3.7 Evaluation

Model performances were compared using the Dice-Sørensen coefficient (Dice) (Dice 1945; Sørensen 1948), which ranges from 0 to 1 with values close to 1 indicating a perfect segmentation overlap between the predicted and ground truth masks. Dice coefficients were computed for each compartment and for each patient scan separately and reported as summary statistics across patients. Specifically, for each segmentation class the mean, standard deviation, and minimum observed Dice coefficients across subjects were considered. Similar statistics were computed for all scans sub-divided by KL-grade classification scores to investigate the effect of OA on model performances.

6.3.8 Statistical Tests

Statistical tests were conducted to assess for differences in the observed mean performance scores on each individual compartment of each dataset (CCBR, OAI & PROOF) between:

1. The MPUnet and KIQ for models trained in the single-cohort setup with limited data (i.e., when the MPUnet and KIQ were trained and evaluated on identical datasets).

2. The MPUnet trained in the single-cohort setup with limited and with additional training data.

3. The MPUnet trained in the single-cohort setup with limited data and the MPUnet trained under the cross-cohort setup.

All reported P-values were computed from paired, two-sided Wilcoxon signed-rank statistics unless explicitly stated otherwise. The Wilcoxon test is non-parametric test and suitable for comparing Dice scores, which are not normally distributed. Performance differences were considered statistically significant at P-value threshold $\alpha = 0.05$. In all cross-validation experiments, each scan in a dataset appears in the test-set of a single fold, and the entire dataset is predicted once and used for computation of evaluation metrics and subsequent statistical tests. In CV the individual hold-out datasets are not statistically independent of each other, because the hold-out data in one fold is in the training data of all other folds. This has to be take into account in when interpreting the statistical results (e.g., see the recent work by Bates et al. 2021).

6.4 Results

6.4.1 Single-Cohort Experiments

Table 6.3 summarizes the segmentation performance of the MPUnet, KIQ and 2D U-Net methods on all three study cohorts (see Table 6.1 and Table 6.2). When trained on the same number of

samples, the MPUnet performed significantly better in terms of the mean macro Dice scores (mean across compartments and patients) on the OAI dataset compared to KIQ (0.86 ± 0.03 vs. 0.84 ± 0.04, $P < 0.05$), the 2D U-Net (0.86 ± 0.03 vs. 0.85 ± 0.03, $P < 0.05$) and the single-view MPUnet (0.86 ± 0.03 vs. 0.85 ± 0.03, $P < 0.05$). The MPunet performed significantly better on the CCBR dataset compared to KIQ (0.83 ± 0.04 vs. 0.81 ± 0.06, $P < 0.05$), the 2D U-Net (0.83 ± 0.04 vs. 0.82 ± 0.05, $P < 0.05$) and the single-view MPunet (0.83 ± 0.04 vs. 0.81 ± 0.04, $P < 0.05$). The MPUnet performed significantly better on the PROOF dataset compared to the 2D U-Net (0.78 ± 0.07 vs. 0.73 ± 0.07, $P < 0.05$) and the single-view MPunet (0.78 ± 0.07 vs. 0.75 ± 0.08, $P < 0.05$) and indifferent from KIQ (0.78 ± 0.07 vs. 0.77 ± 0.07, $P = 0.10$).

Table 6.3 also details the performance of all methods on each individual compartment across the three datasets and shows the minimal Dice scores observed for the compartment across all subjects in the cohort. Across a total of 14 segmentation compartments (tibia bone excluded as it is easily segmented by all methods), the MPUnet performed significantly better than the KIQ model on 11 compartments (TMC, TLC, FMC, FLC, PC, MM and LM on OAI; TMC and FMC on CCBR; FLC and PC on PROOF; $P < 0.05$ for all) and with no significant difference on the remaining 3 (TMC, $P = 0.966$, TLC, $P = 0.170$ and FMC, $P = 0.092$, all on PROOF). The MPUnet performed significantly better than the Paniflov 2D U-Net on 10 compartments (FMC, PC, MM and LM on OAI; FM on CCBR; TMC, TLC, FMC, FLC and PC on PROOF; $P < 0.05$ for all) and with no significant difference on the remaining 4 (TMC, $P = 0.162$, TLC, $P = 0.061$, FLC, $P = 0.087$ on OAI; TMC, $P = 0.182$, on CCBR). The MPUnet performed significantly better than its single-view counterpart on 12 compartments (TMC, FMC, FLC, PC, MM and LM on OAI; TMC and FMC on CCBR; TMC, TLC, FMC and FLC on PROOF; $P < 0.05$ for all) and with no significant difference on the remaining 2 (TLC, $P = 0.056$, on OAI; PC, $P = 0.191$, on PROOF). None of the other models performed significantly better than the MPUnet on any compartment.

Table 6.4 details the performance of each model on the CCBR, OAI and PROOF datasets grouped by Kellgren and Lawrence (KL) grade assessments of each scan. Figure 6.1 shows box-plot Dice score distributions for each compartment of the CCBR dataset as segmented by the MPUnet, KIQ and 2D U-Net models similarly grouped by KL-grades. Similar box-plot figures for the OAI and PROOF datasets are shown in the Appendix Figures B.3 & B.2. On the CCBR dataset, all models had decreasing average performance for increasing KL-grades with mean Dice scores across N=50 KL-0 grade scans and $N = 22$ KL-3 grade scans dropping from 0.84 ± 0.03 to 0.75 ± 0.08 for KIQ, from 0.84 ± 0.03 to 0.76 ± 0.06 for the 2D U-Net, from 0.84 ± 0.02 to 0.73 ± 0.06 for the single-view MPUnet, and from 0.85 ± 0.03 to 0.78 ± 0.06 for the V=6 MPUnet ($P < 0.05$ for all, Mann-Whitney U test). The MPUnet had significantly higher average performance on CCBR KL-3 grade scans compared to both KIQ, 2D U-Net and the single view MPUnet ($P < 0.05$ for all).

On the OAI dataset, for all models there was a non-significant difference between their performances on KL-2 ($N = 10$) and KL-3 ($N = 30$) grade scans (KIQ: 0.84±0.04 and 0.83±0.04 $P = 0.43$; 2D U-Net: 0.85 ± 0.04 and 0.85 ± 0.03, $P = 0.37$; MPUnet ($V = 1$): 0.84 ± 0.03 and 0.85 ± 0.03, $P = 0.30$; MPUnet ($V = 6$): 0.86 ± 0.03 and 0.86 ± 0.04, $P = 0.43$; Mann-Whitney U tests). The MPUnet had significantly higher average performance compared to all other models on the KL-3 group scans with 0.86 ± 0.04 vs. 0.84 ± 0.04 for KIQ, 0.85 ± 0.04 for the 2D U-Net and 0.85 ± 0.03 for the single-view MPUnet ($P < 0.05$ for all). On KL-2 scans the MPUnet performed significantly better than the single-view MPUnet (0.86 ± 0.04 vs 0.84 ± 0.03, $P < 0.05$) and indifferent from both KIQ (0.86 ± 0.04 vs 0.84 ± 0.04, $P = 0.23$) and 2D U-Net (0.86 ± 0.04 vs 0.85 ± 0.04, $P = 0.16$). No statistics were computed for KL-1 or KL-4 scans as the sample sizes of N=2 were too small.

On the PROOF dataset, the MPUnet performed indifferent from KIQ on both $N = 12$ KL-0 scans (0.77 ± 0.07 vs 0.76 ± 0.06, $P = 0.08$) and $N = 11$ KL-1 scans (0.78 ± 0.07 vs 0.77 ± 0.09, $P = 0.41$) and significantly better than the 2D U-Net (0.77 ± 0.07 vs 0.74 ± 0.06, $P < 0.05$) and single-view MPunet (0.77 ± 0.07 vs 0.74 ± 0.08, $P < 0.05$) on KL-0 scans and significantly better than 2D U-Net (0.78 ± 0.07 vs 0.72 ± 0.08, $P < 0.05$) and indifferent from the single-view MPUnet (0.78 ± 0.07 vs 0.76 ± 0.08, $P = 0.07$) on KL-1 scans. No statistics were computed for the N=1 KL-2 or N=1 KL-3 scans as the sample sizes were too small.

Figure 6.2 displays a surface model fit to the manual and MPUnet predicted segmentation masks on a single subject of the OAI cohort. The output was generated by the MPUnet trained in the fixed-split setup (model trained with less data) and having the mean Dice on this image closest to the mean performance over the OAI cohort. Thus, the figure shows the typical performance of the model.

6.4.2 Single-Cohort Experiments: Training with Additional Data

Table 6.3 also summarizes the performance of the MPUnet model when trained on larger versions of the CCBR, OAI and PROOF datasets. On CCBR, the average Dice scores improved slightly from 0.84 ± 0.04 to 0.85 ± 0.04 ($P < 0.05$) on TMC and from 0.82 ± 0.05 to 0.83 ± 0.04 on FMC with the inclusion of additional training data, while the worst-case performance decreased TMC and increased on FMC. On the OAI dataset, the 5-CV models obtained slightly lower Dice scores than the single-split model on average (0.86 ± 0.03 vs. 0.85 ± 0.03). However, for both CCBR and OAI, direct statistical comparisons were not made, because the evaluation datasets differ.

On the PROOF dataset, the addition of 88 OAI scans (significantly different in both resolution, noise level and contrast compared to the scans of PROOF) to the training set significantly improved average Dice scores on FLC (from 0.80 ± 0.07 to 0.83 ± 0.04, $P < 0.05$) and PC (from 0.79 ± 0.07 to 0.81 ± 0.04, $P < 0.05$), non-significantly increased average Dice scores on TLC (from 0.72 ± 0.13 to

0.73 ± 0.13, $P = 0.44$) and FMC (from 0.78 ± 0.08 to 0.79 ± 0.09, $P = 0.09$) and non-significantly decreased performance on TMC (from 0.79 ± 0.06 to 0.78 ± 0.08, $P = 0.58$). The mean macro Dice scores were significantly improved from 0.78 ± 0.07 to 0.79 ± 0.06 ($P < 0.05$).

6.4.3 Cross-Cohort Experiment

Table 6.5 summarizes the performance sores of an MPUnet model trained on images from all the OAI, CCBR and PROOF datasets simultaneously and evaluated on test-set images from OAI and CCBR. The cross-cohort model matched the performance of its specialized counterpart on the CCBR dataset (mean macro Dice scores of 0.83 ± 0.04 and 0.83 ± 0.04, respectively, $P = 0.71$) while the cross-cohort MPUnet model showed significantly decreased, but still high, performance compared to its specialized counterpart on the OAI dataset (0.84 ± 0.04 and 0.86 ± 0.03, respectively, $P < 0.05$).

6.5 Discussion

In this study, three models for automatic MRI knee segmentation were evaluated across three clinical cohorts. Each model was applied as-is without prior tuning of its hyperparameters to simulate a clinical scenario in which the model is to be applied in a new setting (e.g., in a new clinic, for a new scanner or for a new segmentation task), but where manual tuning of the model's hyperparameters is not feasible (e.g., due to lack of technical experts, computational resources or time). The MPUnet was hypothesised to perform well under these restrictions, because it was designed for participation in the 2018 Medical Segmentation Decathlon (Simpson et al. 2019). Participating models were tasked to solve highly variable medical segmentation tasks without (manual) task-specific modifications. The MPUnet ranked 5th without expensive hyperparameter tuning (Simpson et al. 2019) (see Supplementary Information for details). In addition, the MPUnet later scored a top position in the 2019 OA MRI segmentation challenge using the same set of hyperparameters (Arjun Desai et al. 2021).

Here, the MPUnet was compared to the validated KIQ method as well as a state-of-the-art 2D U-Net implementation for knee MRI segmentation by Panfilov et al. (Panfilov et al. 2019) on the OAI, CCBR and PROOF datasets. The considered cohorts varied in both patient demographics, size and scanner sequences, see Tables 6.1 & 6.2. All three models were able to reach high performance on both the CCBR and OAI datasets. However, the MPUnet reached a significantly higher mean macro Dice score on the OAI and CCBR datasets compared to both KIQ and the 2D U-Net. None of the comparison models reached significantly higher Dice scores on any individual compartment across the datasets. The performance scores of the MPUnet were only slightly below the best of models submitted to the 2019 OA MRI segmentation challenge (a set of models which included the MPUnet itself) on the same dataset. There, models achieved mean Dice scores in the range

of 0.86-0.88 but on fewer segmentation compartments than considered in this study (i.e., the 2019 OA MRI segmentation challenge task was simpler) (Arjun Desai et al. 2021). In the challenge, only four compartments were segmented while eight were segmented here. The MPUnet even performed slightly better than the 2D U-Net model, which was tuned specifically for the OAI dataset, and has reported the highest (to our knowledge) mean Dice scores so far with 0.90 ± 0.02 on femoral cartilage, 0.90 ± 0.03 on tibial cartilage, 0.87 ± 0.05 on patellar cartilage and 0.86 ± 0.03 on menisci (Panfilov et al. 2019). It is important to note that the re-trained version of the 2D U-Net model applied to the OAI dataset in this work scores significantly lower on average. Again, this is due to the different number of segmentation compartments considered (four and eight, respectively). For instance, the models trained here must separate the femoral cartilage into both a medial and lateral sub-compartment which is a harder task with uncertainty even in the ground-truth labelling. Interestingly, the KIQ and the MPUnet performed equally good on the small and variable ($N = 25$) PROOF dataset without requiring modifications (for reference, the menisci have previously been automatically segmented with a mean Dice of 0.75 on the same dataset (Xu et al. 2020), but a direct comparison is not possible as the menisci were not segmented here). The 2D U-Net model, however, experienced significant overfitting when trained using its default parameters. Overfitting was decreased by lowering the learning rate and increasing the number of training epochs, but still the obtained 2D U-Net model performed significantly worse than both KIQ and the MPUnet. This result illustrates the premise of this paper, namely that adapting automatic segmentation models in practice is challenging. While the 2D U-Net model of Panfilov et al. is one of the best models fit so far on the OAI dataset for the segmentation of the four considered compartments, that result alone does not provide a guarantee that the model will work well on other, e.g., smaller, datasets or even the same dataset with a different number of segmentation compartments. With systematic hyperparameter tuning, the 2D U-Net model could likely be brought to a high performance also on the PROOF dataset, but such a process may not be feasible in many clinical settings. The KIQ model, although slightly inferior on average to the 2D U-Net on the CCBR and OAI datasets, does not suffer from this limitation when transferred to the small PROOF dataset. This is likely because the framework builds on expert knowledge of knee segmentation, which acts as a strong prior when learning a new dataset. Therefore, the KIQ framework requires less data compared to the 2D U-Net, which must learn from scratch how to segment the 25 new MRIs. Interestingly, the MPUnet, which is also a deep-learning model based on the 2D U-Net and accordingly must also learn from scratch on the small PROOF dataset, did surprisingly well and even outperforms the KIQ framework as measured by the average Dice scores. The MPUnet's robustness and ability to learn from small datasets may result from its unique multi-planar data augmentation strategy. This is supported by the observation that the single-view MPUnet model performs significantly worse than the normal (6

viewed) MPUnet model on all datasets, and often below both KIQ and the 2D U-Net.

The average performance of automatic knee MRI segmentation models are likely to drop for knees of increasing KL-grades as increased OA severity may cause the target compartments to vary abnormally in both shape and volume. Consequently, the robustness and clinical relevance of any automatic model is reflected above all in its performance on high KL-grade scans. This study systematically investigated the performance of all models as a function of KL-grades 0 to 3. The considered cohorts contained too few scans of KL-grade 4 for statistical analysis. On both the OAI and CCBR cohorts, the MPUnet had a significantly higher average performance on knees with moderate OA (KL-3) as compared to all other models. None of the other models performed significantly better than the MPUnet on any individual KL-grade group across the datasets. On the CCBR dataset, all considered models dropped in performance as a function of KL-grade. On the OAI and PROOF datasets, however, the picture is less clear. For instance, the MPUnet performed similar on KL-2 ($N = 10$) and KL-3 ($N = 30$) OAI scans, but with only $N = 2$ KL-1 and $N = 2$ KL-4 grade scans available it could not be concluded if there is an overall decreasing trend or not. Similarly, a decreasing trend could not be concluded on the PROOF data due to the limited number of available scans of KL-grades 2 & 3.

As the performance of deep learning models generally improves with increasing amounts of training data, the potential for further improvement of the MPUnet performance was tested by training separate instances of the model on larger training datasets. As expected, increasing the size of the PROOF training dataset (by using more folds in CV) increased performance as measured by most metrics on all compartments. On the OAI dataset, the performance instead dropped slightly. In both cases, however, a direct comparison is difficult because the evaluation sets differ (evaluation on a fixed test set versus CV). Interestingly, including 88 MRIs from the OAI dataset into the PROOF training set significantly improved the macro Dice performance of the MPUnet. The cross-cohort experiment further showed that a single instance of the MPUnet can learn to segment knee MRIs from two different scanner sequences and patient cohorts with high performance on both. These results suggest that a great potential exists to obtain robust and clinically applicable models by training on larger, merged knee MRI datasets even if they differ with regards to, e.g., scanner sequences, clinical site, and cohort demographics. This strategy of mixing even highly variable training datasets has recently led to the development of robust & clinically applicable models in the field of automated sleep analysis (Perslev, Darkner, et al. 2021). Given the demonstrated high performance of the MPUnet across clinical cohorts, MRI sequences and KL-grades, such a model, if trained on enough and variable data, is perhaps archivable also for knee MRI segmentation and could ultimately serve as a ready-to-use, robust model for general knee MRI segmentation.

The pre-trained MPUnet models are made available. These models may be used directly or serve

as initializations for training new models. This transfer learning can help building well generalizing models for new data even if the new dataset is very small.

6.5.1 Limitations

This study considered mean Dice scores as a direct proxy for general knee MRI segmentation performance. Further studies should be made to address if the presented observations hold also for other clinically relevant metrics such as surface distances, volumes, etc. Ultimately, future studies should address if the segmentation masks obtained by deep learning allow for accurate assessments of pathologies such as OA associated cartilages. In addition, this study did not include data from all major producers of MRI scanners (e.g., GE Healthcare). Finally, it is a limitation of the study that data selection was done retrospectively.

6.6 Conclusion

This study found that the MPUnet improves on the state-of-the-art in knee MRI segmentations across cohorts without the need for manual adaptations. It was found accurate even on high KL-grade scans and could learn across multiple cohorts at once. This robustness of the MPUnet makes it practical and applicable also for research groups with limited specialist knowledge of deep learning, because the framework may be easily adapted to new data or even applied directly using one of the pre-trained models that were made available.

Data Availability

Pre-trained MPUnet models for cohorts OAI, CCBR and PROOF are available at `https://sid.erda.dk/cgi-sid/ls.py?share_id=DQADRdWlID`. The multi-planar convolutional neural network method is available as open-source software. The software can be used without prior knowledge of deep learning, is open sourced under the MIT license and is available along with tutorials at `https://github.com/perslev/MultiPlanarUNet`. To fit the model to new MRI sequences, a set of manually annotated segmentation masks are required. With these at hand, the included Python scripts will perform training, evaluation, and predictions on future images with launching the script on properly organized data folders being the only involved human action.

The software requires just a single GPU but can utilize additional GPUs if available. For most applications, 12GB GPU memory is required for optimal performance. On our system with a single GPU segmenting a new scan takes 2-6 minutes.

Acknowledgements

Mathias Perslev and Christian Igel gratefully acknowledge support from the Independent Research Fund Denmark through the project "U-Sleep" (project number 9131-00099B). Akshay Pai and Christian Igel gratefully acknowledge support from The Danish Industry Foundation as part of the initiative AI Denmark. The OAI collection was provided by the OAI. The OAI is a public–private partnership comprised of five contracts (N01-AR-2-2258, N01-AR-2-2259, N01-AR-2-2260, N01-AR-2-2261, and N01- AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health. This manuscript was prepared using an OAI public use dataset and does not necessarily reflect the opinions or views of the OAI investigators, the NIH, or the private funding partners.

Table 6.3: Single-Cohort Experiments: Segmentation performance across subjects for the MPUnet, single-view MPUnet, 2D U-Net and KIQ methods on the OAI, CCBR and PROOF cohorts. Individual scores where the other models score better than the MPUnet are marked in red. Accuracy is given as the Dice volume overlap showing mean ± std and minimum values. P-values for the paired, two-sided Wilcoxon signed-rank statistic are shown for all compartments comparing the MPUnet performance against itself when trained on additional data, the KIQ method, the single-view MPUnet and the 2D U-Net when evaluated on identical dataset. CV = Cross Validation. LOO = Leave One Out (number of CV folds identical to the number of evaluation images). [a]Tibia bone excluded from computation of Macro Dice scores. [b]Lower LR, higher epochs compared to paper.

Dataset	Method	Eval. type	Eval. images	Tibia bone[a]	Tibial medial cartilage	Tibial lateral cartilage	Femoral medial cartilage	Femoral lateral cartilage	Patellar cartilage	Medial meniscus	Lateral meniscus	Macro Dice
CCBR	KIQ	Fixed split	110	-	0.83 ± 0.06 0.47, $P < 0.05$	-	0.79 ± 0.06 0.52, $P < 0.05$	-	-	-	-	0.81 ± 0.06 0.57, $P < 0.05$
	Panfilov 2D U-Net	Fixed split	110	-	0.83 ± 0.06 0.57, $P = 0.18$	-	0.81 ± 0.05 0.64, $P < 0.05$	-	-	-	-	0.82 ± 0.05 0.64, $P < 0.05$
	MP ($V = 1$)	Fixed split	110	-	0.82 ± 0.06 0.60, $P < 0.05$	-	0.80 ± 0.06 0.57, $P < 0.05$	-	-	-	-	0.81 ± 0.06 0.59, $P < 0.05$
	MP ($V = 6$)	Fixed split	110	-	0.84 ± 0.04 0.68	-	0.82 ± 0.05 0.654	-	-	-	-	0.83 ± 0.04 0.69
	MP ($V = 6$)	5-CV	140	-	0.85 ± 0.04 0.65	-	0.83 ± 0.04 0.68	-	-	-	-	0.84 ± 0.04 0.68
OAI	KIQ	Fixed split	44	0.98 ± 0.00 0.98, $P < 0.05$	0.84 ± 0.05 0.69, $P < 0.05$	0.89 ± 0.04 0.73, $P < 0.05$	0.83 ± 0.05 0.68, $P < 0.05$	0.86 ± 0.04 0.73, $P < 0.05$	0.78 ± 0.11 0.40, $P < 0.05$	0.80 ± 0.10 0.34, $P < 0.05$	0.86 ± 0.04 0.75, $P < 0.05$	0.84 ± 0.04 0.72, $P < 0.05$
	Panfilov 2D U-Net	Fixed split	44	0.89 ± 0.01 0.87, $P < 0.05$	0.85 ± 0.05 0.71, $P = 0.16$	0.89 ± 0.03 0.80, $P = 0.06$	0.85 ± 0.05 0.65, $P < 0.05$	0.88 ± 0.04 0.74, $P = 0.08$	0.81 ± 0.12 0.33, $P < 0.05$	0.82 ± 0.07 0.57, $P < 0.05$	0.87 ± 0.03 0.79, $P < 0.05$	0.85 ± 0.03 0.77, $P < 0.05$
	MP ($V = 1$)	Fixed split	44	0.98 ± 0.0 0.98, $P < 0.05$	0.837 ± 0.052 0.68, $P < 0.05$	0.89 ± 0.04 0.75, $P = 0.06$	0.84 ± 0.05 0.61, $P < 0.05$	0.86 ± 0.05 0.70, $P < 0.05$	0.82 ± 0.10 0.51, $P < 0.05$	0.82 ± 0.07 0.60, $P < 0.05$	0.88 ± 0.04 0.74, $P < 0.05$	0.85 ± 0.03 0.76, $P < 0.05$
	MP ($V = 6$)	Fixed split	44	0.98 ± 0.0 0.98	0.85 ± 0.05 0.717	0.90 ± 0.04 0.79	0.86 ± 0.05 0.66	0.88 ± 0.04 0.73	0.83 ± 0.11 0.26	0.83 ± 0.06 0.66	0.89 ± 0.03 0.82	0.86 ± 0.03 0.75
	MP ($V = 6$)	5-CV	176 (174)	-	0.845 ± 0.049 0.67	0.89 ± 0.03 0.76	0.88 ± 0.03 0.71		0.81 ± 0.10 0.26	0.82 ± 0.07 0.55	0.87 ± 0.03 0.66	0.85 ± 0.03 0.70
PROOF	KIQ	25-CV	25	0.96 ± 0.02 0.91, $P < 0.05$	0.79 ± 0.06 0.61, $P = 0.97$	0.76 ± 0.09 0.41, $P = 0.17$	0.77 ± 0.10 0.44, $P = 0.09$	0.80 ± 0.05 0.64, $P < 0.05$	0.72 ± 0.11 0.36, $P < 0.05$	-	-	0.77 ± 0.07 0.52, $P = 0.10$
	Panfilov 2D U-Net[b]	25-CV	25	0.97 ± 0.01 0.94, $P < 0.05$	0.73 ± 0.09 0.48, $P < 0.05$	0.67 ± 0.11 0.39, $P < 0.05$	0.73 ± 0.08 0.52, $P < 0.05$	0.75 ± 0.07 0.59, $P < 0.05$	0.76 ± 0.07 0.60, $P < 0.05$	-	-	0.73 ± 0.07 0.54, $P < 0.05$
	MP ($V = 1$)	25-CV	25	0.95 ± 0.05 0.75, $P < 0.05$	0.76 ± 0.09 0.41, $P < 0.05$	0.69 ± 0.15 0.20, $P < 0.05$	0.75 ± 0.09 0.48, $P < 0.05$	0.77 ± 0.09 0.38, $P < 0.05$	0.78 ± 0.06 0.60, $P = 0.19$	-	-	0.75 ± 0.08 0.50, $P < 0.05$
	MP ($V = 6$)	25-CV	25	0.96 ± 0.02 0.89	0.79 ± 0.06 0.63	0.72 ± 0.13 0.29	0.78 ± 0.08 0.50	0.80 ± 0.07 0.47	0.79 ± 0.07 0.59	-	-	0.78 ± 0.07 0.56
	MP ($V = 6$)	25-CV + 88 OAI	25	-	0.78 ± 0.08 0.53, $P = 0.58$	0.73 ± 0.13 0.26, $P = 0.44$	0.79 ± 0.09 0.45, $P = 0.09$	0.83 ± 0.04 0.67, $P < 0.05$	0.81 ± 0.04 0.70, $P < 0.05$	-	-	0.79 ± 0.06 0.60, $P < 0.05$

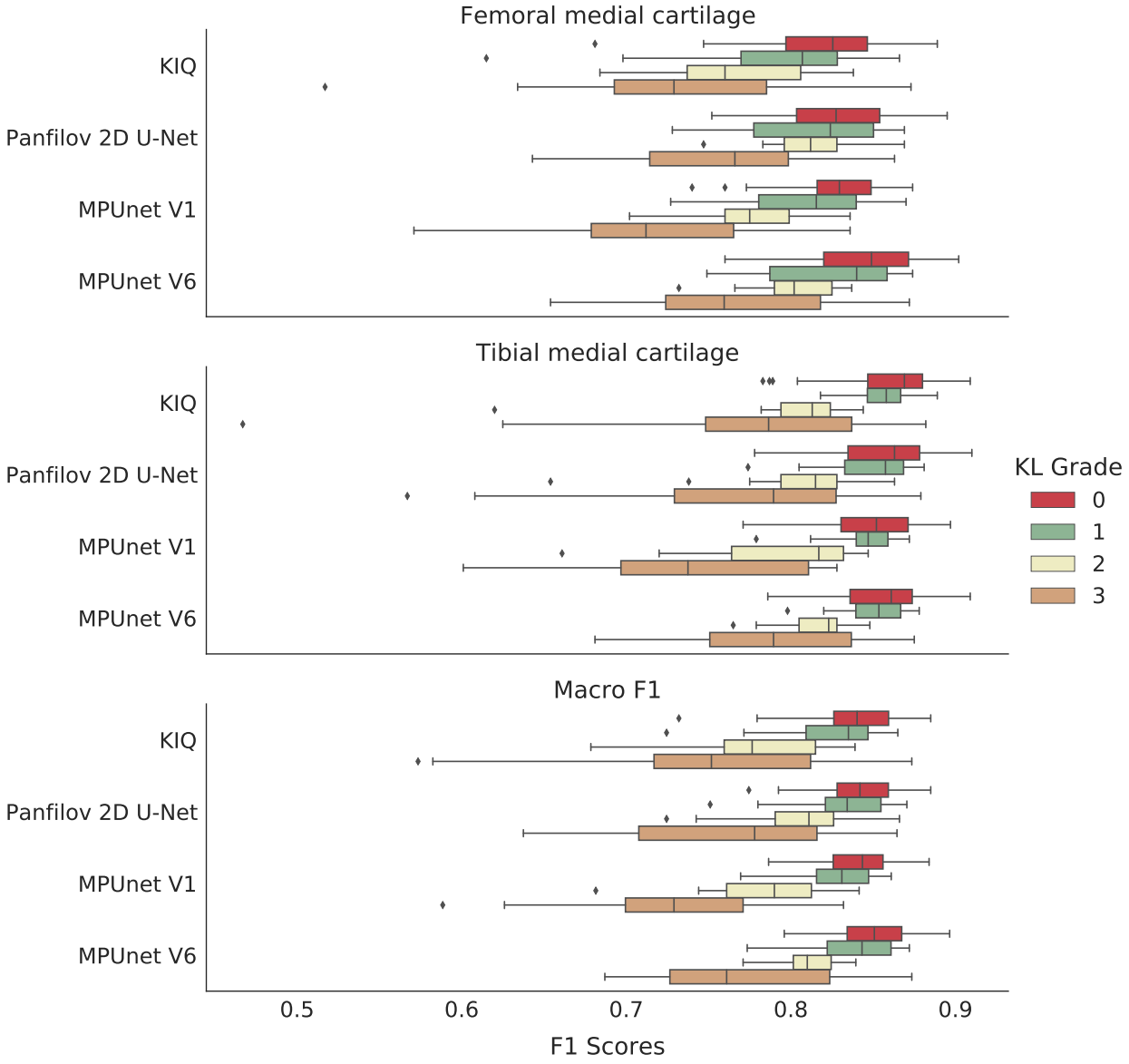


Figure 6.1: Box-plots showing the distribution of Dice scores for the MPUnet, KIQ and the 2D U-Net on the CCBR dataset grouped according to the KL-grade score of the individual MRIs. (a) Dice scores on the Femoral Medial Cartilage. (b) Dice scores on the Tibial Medial Cartilage. (c) Macro Dice scores.

Table 6.4: Single-Cohort Experiments – KL Groups: Segmentation performance across subjects for the MPUnet, single-view MPUnet, 2D U-Net and KIQ methods on the OAI, CCBR and PROOF cohorts on KL sub-groups. Individual scores where the other models score better than the MPUnet are marked in red. [a]Lower LR, higher epochs compared to Panfilov et al., 2019 (13).

Dataset	Method	Eval. type	Eval. images	KL 0	KL 1	KL 2	KL 3	KL 4
CCBR	KIQ	Fixed split	50 / 24 / 13 / 22 / 0	0.84 ± 0.03	0.82 ± 0.03	0.78 ± 0.04	0.75 ± 0.08	-
				0.73, $P < 0.05$	0.72, $P < 0.05$	0.68, $P < 0.05$	0.57, $P < 0.05$	
	Panfilov 2D U-Net	Fixed split	50 / 24 / 13 / 22 / 0	0.84 ± 0.03	0.83 ± 0.03	0.80 ± 0.04	0.76 ± 0.06	-
				0.774, $P < 0.05$	0.75, $P = 0.03$	0.72, $P = 0.74$	0.64, $P < 0.05$	
	MP ($V = 1$)	Fixed split	50 / 24 / 13 / 22 / 0	0.84 ± 0.02	0.83 ± 0.03	0.78 ± 0.04	0.73 ± 0.06	-
				0.79, $P < 0.05$	0.77, $P < 0.05$	0.68, $P < 0.05$	0.59, $P < 0.05$	
	MP ($V = 6$)	Fixed split	50 / 24 / 13 / 22 / 0	0.85 ± 0.03, 0.80	0.84 ± 0.03, 0.77	0.81 ± 0.02, 0.77	0.78 ± 0.06, 0.69	-
OAI	KIQ	Fixed split	0 / 2 / 10 / 30 / 2	-	0.88 ± 0.03	0.84 ± 0.04	0.83 ± 0.04	0.83 ± 0.02
					0.86, $P = \text{n/a}$	0.76, $P = 0.23$	0.72, $P < 0.05$	0.82, $P = \text{n/a}$
	Panfilov 2D U-Net	Fixed split	0 / 2 / 10 / 30 / 2	-	0.87 ± 0.03	0.85 ± 0.04	0.85 ± 0.03	0.86 ± 0.02
					0.85, $P = \text{n/a}$	0.78, $P = 0.16$	0.77, $P < 0.05$	0.85, $P = \text{n/a}$
	MP ($V = 1$)	Fixed split	0 / 2 / 10 / 30 / 2	-	0.87 ± 0.02	0.84 ± 0.03	0.85 ± 0.03	0.86 ± 0.01
					0.85, $P = \text{n/a}$	0.77, $P < 0.05$	0.76, $P < 0.05$	0.85, $P = \text{n/a}$
	MP ($V = 6$)	Fixed split	0 / 2 / 10 / 30 / 2	-	0.88 ± 0.03	0.86 ± 0.03	0.86 ± 0.04	0.87 ± 0.01
					0.86	0.78	0.75	0.87
PROOF	KIQ	25-CV	12 / 11 / 1 / 1 / 0	0.76 ± 0.06	0.77 ± 0.09	0.81 ± 0.00	0.80 ± 0.00	-
				0.63, $P = 0.08$	0.52, $P = 0.41$	0.81, $P = \text{n/a}$	0.80, $P = \text{n/a}$	
	Panfilov 2D U-Net[b]	25-CV	12 / 11 / 1 / 1 / 0	0.74 ± 0.06	0.72 ± 0.08	0.76 ± 0.00	0.71 ± 0.00	-
				$0.60, P < 0.05$	$0.54, P < 0.05$	$0.76, P = \text{n/a}$	$0.71, P = \text{n/a}$	
	MP ($V = 1$)	25-CV	12 / 11 / 1 / 1 / 0	0.74 ± 0.08	0.76 ± 0.08	0.77 ± 0.00	0.77 ± 0.00	-
				0.50, $P < 0.05$	0.52, $P = 0.07$	0.77, $P = \text{n/a}$	0.77, $P = \text{n/a}$	
	MP ($V = 6$)	25-CV	12 / 11 / 1 / 1 / 0	0.77 ± 0.07	0.78 ± 0.07	0.82 ± 0.00	0.79 ± 0.00	-
				0.56	0.59	0.82	0.79	

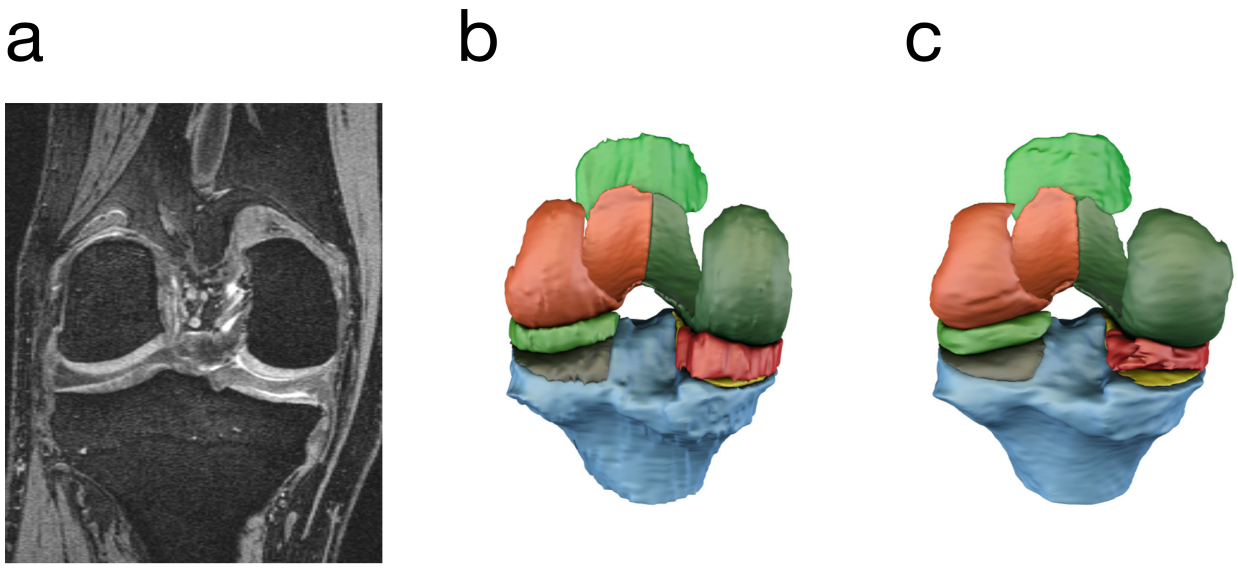


Figure 6.2: Surface models visually comparing the expert annotated segmentation (b) to the annotations of MPUnet (c) on an average performing sample of the OAI dataset. (a) Shows a reference coronal slice from the MRI volume with KL-grade = 3.

Table 6.5: Cross-Cohort Experiment: Segmentation performance across subjects in the test-splits of OAI and CCBR of a single MPUnet model instance trained on MRIs from all of the CCBR, OAI and PROOF cohorts. Accuracy is given as the Dice volume overlap showing mean±std and minimum values. P-values compare the per-compartment mean Dice scores of the cross-cohort model to the MPUnet trained and evaluated on the individual cohorts.

Method	MPUnet ($V = 6$)	
Training images	30 CCBR + 44 OAI + 25 PROOF	
Evaluation images	110 CCBR	44 OAI
Tibial medial cartilage	0.84 ± 0.05	0.83 ± 0.06
	0.59, $P = 0.49$	0.59, $P < 0.05$
Femoral medial cartilage	0.82 ± 0.05	0.85 ± 0.05
	0.68, $P = 0.07$	0.66, $P < 0.05$
Macro Dice	0.83 ± 0.04	0.84 ± 0.04
	0.66, $P = 0.71$	0.72, $P < 0.05$

Chapter 7

Related work

The following papers were all published and co-authored during the PhD but not enclosed in this thesis because they contain tangentially related work or challenge participation papers written together with large numbers of authors.

7.1 The Medical Segmentation Decathlon

This paper by Simpson et al. (2019) presents the organization, outcomes, and conclusions of the 2018 MICCAI Medical Segmentation Decathlon challenge. We participated in this challenge using the MPUNet model described in Paper A. As briefly mentioned in Chapter 3, Section 3.2.4, the MSD challenge encouraged teams to develop machine learning pipelines capable of automatically adapting themselves to learn models for ten distinct medical image segmentation tasks involving MRI and CT scans. A labelled training dataset was provided for each task, but no task-specific information could be given to the system, and manual hyperparameter tuning was not permitted. As a result, the proposed methods must be fully automated, using only a labelled dataset as input to generate a segmentation model for a specific task.

Several innovative pipelines were developed during the challenge. The winning team created the now widely adopted nnU-Net framework (Isensee et al. 2018). This framework combines automatic hyperparameter specification using a set of custom heuristics that analyze or *fingerprint* the provided training dataset, along with an automatic selection of an appropriate U-Net-type model architecture from a collection of three options, determined through cross-validation experiments. The model candidates include a 2D U-Net, a 3D U-Net, and a cascaded 2D U-Net, which generates an initial low-resolution segmentation, subsequently refined by a 3D U-Net model to produce the final high-resolution segmentation. An instance of each candidate model is trained, and the single model or ensemble combining the outputs of two individual models and achieving the best F1 score in a 5-fold cross-validation experiment is used to predict the test images. The nnU-Net framework ranked first in 9 out of 10 tasks, demonstrating versatility. However, one drawback of this approach is the computational overhead of training multiple candidate models during cross-validation experiments.

The second-place team adopted a distinct approach with some similarities to our MPUNet method of Paper A. Utilizing a single, relatively constrained model architecture. They focused on designing an optimization process that maximizes the use of available 3D information from medical volumes

while training parameter-efficient models that process only 2D inputs or thin 3D patches (i.e., a stack of a small number of 2D image planes sampled along the third image axis) (Xia et al. 2020). Employing a data-resampling strategy similar to our MPUNet method, they utilized that multiple 2D views could be generated from a 3D volume by applying rotation transformations. These views can either be fed into a single 2D model, as in our method, to expand the available training dataset, or into separate models specialized for each view orientation, as in their approach. Additionally, they developed a semi-supervised optimization strategy based on co-training (Blum et al. 1998) that enabled their method to incorporate information from labelled and unlabeled data during training. Their models were initialized with weights obtained from the only tangentially related tasks on natural images.

In contrast to approaches like the nnU-Net, which selects suitable model architectures through heuristics and automated experimentation, the methods of (Xia et al. 2020) and our MPUNet aim to make the machine learning pipeline robust across tasks by employing a single model architecture with high approximation capacity that can be trained on small datasets with limited overfitting because it operates with high statistical efficiency in 2D (or thin 3D patches) while the labelled data is 3D. However, it is essential to note that these multi-view approaches and the model selection strategy of the nnU-Net are orthogonal methods, operating at the data and model levels, respectively, and could be implemented together. For example, multi-planar data augmentation could be incorporated within the nnU-Net to enhance the performance of the 2D candidate model. The nnU-Net framework would still apply its hyperparameter selection heuristics and automatic model selection techniques to choose and configure an optimal model for a given task.

The challenge also highlighted a set of best practices that were implemented by most teams. For example, data augmentation techniques were employed by all top-performing groups. Data augmentation involves artificially expanding the available training dataset size by applying one or more (often randomized) image transformation functions to each training example, generating subtle variations in the observed data. For instance, large sets of randomly defined affine transformations can be applied to all available training images to introduce model invariance to such transformations (refer to Shorten et al. (2019) for a review of data augmentation techniques). Additionally, most teams implemented specialized loss functions or data-resampling procedures to address label imbalance issues, which are often prominent in medical images where the foreground object of interest is considerably smaller than the remaining image volume.

Since the original challenge, several new submissions have been made, some of which have surpassed the performance of the initial nnU-Net challenge winners. Part of this improvement can be attributed to the new submissions having the advantage of tuning their systems on all ten datasets, unlike the original challenge participants who only had access to seven datasets during the open

phase. However, some methodological advances have also been made since the 2018 challenge. For example, several newer methods rely on auto-ML-inspired techniques (refer to Hutter et al. (2019) and X. He et al. (2021) for reviews), specifically neural architecture search (NAS, see Elsken et al. (2019) for a review). In NAS, a neural network architecture is automatically generated by iteratively selecting new candidate models from a defined search space using, for instance, evolutionary optimization algorithms (Real et al. 2017) or reinforcement learning (Zoph et al. 2016). However, while NAS and other auto-ML techniques may enhance a system's overall performance, they often come with significant computational overhead, as numerous candidate models are evaluated through optimization. NAS was employed by Y. He et al. (2021), who led the challenge until recently when Hatamizadeh, Nath, et al. (2022) took the lead, incorporating results from the more recent Vision Transformer architectures of UNETR (Hatamizadeh, Tang, et al. 2022) and Swin Transformers (Ze Liu et al. 2021). These newer Transformer architectures aim to combine the strengths of FCN models like the U-Net (e.g., hierarchical feature encoding and encoder-decoder bottleneck structures) with the advantages of Transformers (e.g., their ability to process long-range dependencies).

Many of the techniques mentioned operate at different levels of the machine learning pipeline and may be composable. Based on the findings of the MSD, one could hypothesize that a promising approach to achieving highly generalizable machine learning pipelines for medical segmentation – provided computational overhead is not a significant constraint – involves a combination of an automatically configured model architecture (either through NAS or similar auto-ML techniques or by selecting a candidate from a set of models). This architecture would first be pre-trained on a large and diverse collection of unlabelled images or labelled images for some related task and then fine-tuned using extensive data augmentation (e.g., multi-view training and classical augmentations) on a smaller set of labelled image examples for the specific target task.

The conclusions from the 2018 MSD challenge will be further explored concerning the general findings of this thesis in the Discussion chapter 13.

7.2 The Liver Tumor Segmentation Benchmark (LiTS)

This paper by Bilic et al. (2023) presents the setup, results, and conclusions of the Liver Tumor Segmentation Benchmark (LiTS) challenge. The challenge was hosted in three editions, first at a workshop during the 2017 IEEE International Symposium on Biomedical Imaging (ISBI), then at the 2017 MICCAI conference, and finally as the liver segmentation task of the 2018 MSD MICCAI challenge described earlier. Participants were invited to develop segmentation models for segmenting the liver and primary and secondary liver tumours in CT images. We participated in this challenge through the 2018 MSD using the MPUNet model detailed in Paper A.

The dataset comprised 131 training and 70 testing images from eight geographically dispersed clinical institutions. The imaged tumours exhibited considerable variability, encompassing primary and secondary tumours with diverse shapes, locations, sizes, contrasts, and densities. As highlighted in the Motivation chapter 1, liver tumours are complex segmentation targets due to these factors, even for human expert annotators.

The challenge demonstrated that many automatic segmentation systems also struggle with these technical difficulties. While all methods accurately segmented the liver (average per-subject F1/Dice score overlaps of 0.92-0.96 for most teams across the three editions), the liver tumours were segmented with an average F1 overlap in the range of 0.64 – 0.67 in the 2017 edition. However, the top-performing methods in the 2017 (MICCAI) and 2018 (MICCAI MSD) editions significantly improved these results, achieving average F1 overlaps of 0.70 and 0.74, respectively. The latter score was achieved by the nnU-net, the winner of the 2018 MSD challenge (Isensee et al. 2018). These results demonstrate that a task-agnostic system can outperform specialized systems from the two editions where liver tumour segmentation was the sole target task. However, most submissions to the 2018 MSD challenge did not accomplish this feat. The median score on the tumour class among the 19 participants was 0.54, below the typical performance of systems submitted to the 2017 ISBI edition. For example, in our submission, the MPUNet segmented the liver tumour class with a mean F1 score of 0.57, corresponding to the 8^{th} highest of the 19 participants. It's overall ranking in the challenge was $5^{\text{th}}/6^{\text{th}}$, indicating a worse-than-usual performance on the liver tumour segmentation task. These results will further be discussed in the Discussion chapter 13.

Over the three editions of the challenge, a total of 73 fully automated segmentation models were submitted. Similar to the 2018 MSD challenge, the U-Net architecture was employed in most systems, often in a cascaded setup where separate models first performed coarse-grained outlining followed by fine-detailed segmentation. Almost all teams utilized various types of data augmentation. Additionally, many teams implemented post-processing of the segmentation masks, such as discarding detected tumours outside the segmented liver region, filling holes in tumours, and more.

7.3 The International Workshop on Osteoarthritis Imaging Knee MRI Segmentation Challenge

This paper by Arjun Desai et al. (2021) presents the setup, results, and conclusions of the 2019 International Workshop on Osteoarthritis Imaging (IWOI) Knee MRI Segmentation Challenge. The challenge invited participants to develop segmentation models for segmenting knee cartilages in MRI to assess the clinical efficacy of using automatic models to extract osteoarthritis (OA) biomarkers. We participated in this challenge using the MPUNet model described in Paper A without further

modifications, aiming to test the model's ability to solve a specific task compared to submissions explicitly tuned for the challenge.

A total of five teams participated in the challenge, with one team submitting an entry after the challenge concluded. Two teams made two submissions each. All eight submissions utilized U-Net-inspired FCN models but with varying numbers of parameters, optimization hyperparameters, and loss functions. The segmentations were assessed using standard overlap metrics, such as the F1 score. Additionally, the error in estimating cartilage thickness, a potential biomarker for OA progression assessment, was evaluated by comparing manually defined and automatic segmentations. All methods achieved accurate segmentations, with average F1 scores ranging from 0.81 to 0.90 for all four targets (femoral, tibial, patellar cartilages, and the meniscus). The top four performing teams, including the MPUNet, demonstrated no significant differences in F1 overlap for any targets. The cartilage thickness error was also similar among these four teams, with low mean percentage errors between 4,% and 6,% for the femoral and tibial cartilage targets (within the range of practical limitations defined by the scanning resolution). However, the patellar cartilage showed more significant and varied errors, with mean errors between 6,% and 10,%. The thickness estimates were inconsistent across scans for all networks, potentially limiting their clinical applicability. The MPUNet demonstrated average thickness errors of 4,%, 6,%, and 6,% for the three cartilage targets.

Despite achieving high overall segmentation overlap scores, the challenge revealed that all networks performed worse in cartilage areas most commonly affected by OA. These findings suggest that various networks (although all U-Net-inspired) optimized using different hyperparameters and loss functions may produce similar knee cartilage segmentations in MRI. The limitations of these networks are likely not due to their architecture or optimization hyperparameters but rather the lack of training data that adequately represents the clinical variability of cartilages impacted by OA.

The challenge also made an important observation that there was only a weak correlation between F1 segmentation overlap scores and cartilage thickness estimation errors. This finding indicates that the effectiveness of simple overlap metrics, such as F1, is limited in assessing the clinical applicability of different models that all achieve high F1 scores. Slight improvements to the F1 score may not significantly impact the accuracy of the clinical endpoint variable of interest. The limitations of evaluating clinical segmentation models using simple F1 scores are further discussed in the Limitations chapter 13.4 in Part IV below.

7.4 Towards Automatic Cartilage Quantification in Clinical Trials

This paper by Dam, Arjun Desai, et al. (2023) presents the setup, results, and conclusions of a follow-up challenge to the IWOI Knee MRI Segmentation Challenge by Arjun Desai et al. (2021), discussed

above. All six teams that participated in the original challenge were invited to join the follow-up challenge, and they all accepted. Each team used their original model, trained on the dataset from the initial challenge, to segment a total of 1,130 new knee MRIs from a different sub-cohort within the same study (the Osteoarthritis Initiative (OAI) multi-visit cohort, Nevitt et al. 2006) from which the data for the original challenge was selected. The sub-cohort consisted of 556 subjects imaged during baseline and follow-up visits one year later.

The challenge aimed to assess if methods developed for the 2019 IWOAI challenge could reliably quantify OA imaging biomarkers, such as cartilage loss, in a larger cohort with manually measured cartilage volume scores. The challenge focused on a longitudinal trial where scanner changes and drifts caused by software updates or part replacements could alter MRI appearances between a subject's visits, even when using the same scanner and sequence throughout the study.

The methods were primarily assessed based on their ability to detect cartilage volume changes between the baseline and follow-up visits. As observed in the original challenge, all methods demonstrated high segmentation accuracy, with correlations between manual and automatic cartilage volume measurements ranging from 0.82 to 0.95. The automatic methods tended to either overestimate or underestimate cartilage volumes but were relatively consistent in their tendencies. The challenge concluded that both automatic and manual segmentations had similar sensitivity to cartilage volume changes, at least for those compartments where the ground truth was reliable (see Dam, Arjun Desai, et al. (2023) for further details).

The highest sensitivity to cartilage volume change was achieved by a slightly modified version of our MPUNet (in which the typical ReLU activation functions were replaced with ELU activation functions) specifically for the lateral tibial cartilage compartment. In this compartment, the MPUNet demonstrated sensitivity to detecting volume changes at least as well as manual annotations, suggesting that this method and other high-performing submissions from the challenge may be used in future clinical studies to quantify OA imaging biomarkers. However, it is essential to note that all automatic methods were more sensitive to scanner drift and shift events than manual annotations. Therefore, continuous analysis of the potential impact of such events on the obtained automatic quantifications is essential throughout the study. To maintain optimal performance, models may need to be re-trained or fine-tuned on newly annotated scans whenever significant scanner drift is detected.

7.5 Accurate Segmentation of Dental Panoramic Radiographs with U-Nets

This paper describes concurrent work to Paper A on the MPUnet (T. Koch et al. 2019). It outlines several best practices for applying U-Net-like FCN models to medical image segmentation tasks, some of which were also implemented in Paper A. Specifically, a U-Net architecture was designed for segmenting dental radiographs. A series of experiments were conducted to explore the potential for optimization using various loss functions, input image patching strategies (since processing the large radiograph images as a whole was not feasible at the time, even for FCNs), data augmentation techniques(reflections of the input radiographs to utilize their natural symmetry properties), ensemble predictions generated by multiple independently trained networks, and a method known as *test-time* augmentation. This last approach involved predicting the same radiographs using the same network multiple times after applying different transformation functions.

The concept of test-time augmentation was already introduced in the AlexNet paper by Krizhevsky et al. (2017). Their model was applied to multiple overlapping patches extracted from the input to produce a more accurate average segmentation. Test-time augmentation inspired the multi-planar data augmentation scheme presented in Paper A, where the training data augmentation scheme introduces model rotation equivariance properties which can be further utilized at test time to predict image volumes as seen from multiple view orientations.

Part III

Sleep Staging

Chapter 8

Paper C

U-Time: A fully convolutional network for time series segmentation applied to sleep staging

Authors

Mathias Perslev[a], Michael Hejselbak Jensen[a], Sune Darkner[a], Poul Jørgen Jennum[b] and Christian Igel[a].

[a]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

[b]Danish Center for Sleep Medicine, Rigshospitalet, Glostrup, Denmark

Published

Advances in Neural Information Processing Systems (NeurIPS 2019), pp. 4417-4428, 2019. URL: `https://proceedings.neurips.cc/paper/2019/file/57bafb2c2dfeefba931bb03a835b1fa9-Paper.pdf`

8.1 Abstract

Neural networks are becoming more and more popular for the analysis of physiological time-series. The most successful deep learning systems in this domain combine convolutional and recurrent layers to extract useful features to model temporal relations. Unfortunately, these recurrent models are difficult to tune and optimize. In our experience, they often require task-specific modifications, which makes them challenging to use for non-experts. We propose U-Time, a fully feed-forward deep learning approach to physiological time series segmentation developed for the analysis of sleep data. U-Time is a temporal fully convolutional network based on the U-Net architecture that was originally proposed for image segmentation. U-Time maps sequential inputs of arbitrary length to sequences of class labels on a freely chosen temporal scale. This is done by implicitly classifying every individual time-point of the input signal and aggregating these classifications over fixed intervals to form the final predictions. We evaluated U-Time for sleep stage classification on a large collection of sleep electroencephalography (EEG) datasets. In all cases, we found that U-Time reaches or outperforms current state-of-the-art deep learning models while being much more robust in the training process and without requiring architecture or hyperparameter adaptation across tasks.

8.2 Introduction

During sleep our brain goes through a series of changes between different *sleep stages*, which are characterized by specific brain and body activity patterns (Iber et al. 2007; Kales et al. 1968). *Sleep staging* refers to the process of mapping these transitions over a night of sleep. This is of fundamental importance in sleep medicine, because the sleep patterns combined with other variables provide the basis for diagnosing many sleep related disorders (Sateia 2014). The stages can be determined by measuring the neuronal activity in the cerebral cortex (via electroencephalography, EEG), eye movements (via electrooculography, EOG), and/or the activity of facial muscles (via electromyography, EMG) in a *polysomnography* (PSG) study (see Figure C.1 in the Supplementary Material). The classification into stages is done manually. This is a difficult and time-consuming process, in which expert clinicians inspect and segment the typically 8–24 hours long multi-channel signals. Contiguous, fixed-length intervals of 30 seconds are considered, and each of these *segments* is classified individually.

Algorithmic sleep staging aims at automating this process. Recent work shows that such systems can be highly robust (even compared to human performance) and may play an important role in developing novel biomarkers for sleep disorders and other (e.g., neurodegenerative and psychiatric) diseases (Schenck et al. 2014; Stephansen et al. 2018; Warby et al. 2014). Deep learning is becoming increasingly popular for the analysis of physiological time-series (Faust, Hagiwara, et al. 2018) and has already been applied to sleep staging (Faust, Razaghi, et al. 2019; Robert et al. 1998; Ronzhina et al. 2012). Today's best systems are based on a combination of convolutional and recurrent layers (Biswal, Kulas, et al. 2017; Supratak et al. 2017). While recurrent neural networks are conceptually appealing for time series analysis, they are often difficult to tune and optimize in practice, and it has been found that for many tasks across domains recurrent models can be replaced by feed-forward systems without sacrificing accuracy (Bai, J. Z. Kolter, et al. 2018; Q. Chen et al. 2017; Vaswani et al. 2017).

This study introduces U-Time, a feed-forward neural network for sleep staging. U-Time as opposed to recurrent architectures can be directly applied across datasets of significant variability without any architecture or hyperparameter tuning. The task of segmenting the time series is treated similar to image segmentation by the popular U-Net architecture (Ronneberger et al. 2015). This allows segmentation of an entire PSG in a single forward pass and to output sleep stages at any temporal resolution. Fixing a temporal embedding, which is a common argument against feed-forward approaches to time series analysis, is no problem, because in our setting the full time series is available at once and is processed entirely (or in large chunks) at different scales by the special network architecture.

In the following, we present our general approach to classifying fixed length continuous segments of physiological time series. In Section 8.4, we apply it to sleep stage classification and evaluate it on 7 different PSG datasets using a fixed architecture and hyperparameter set. In addition, we performed many experiments with a state-of-the-art recurrent architecture, trying to improve its performance over U-Time and to assess its robustness against architecture and hyperparameter changes. These experiments are listed in the Supplementary Material. Section 8.5 summarizes our main findings, before we conclude in Section 8.6.

8.3 Method

U-Time is a fully convolutional encoder-decoder network. It is inspired by the popular U-Net architecture originally proposed for image segmentation (T. Koch et al. 2019; Perslev, Dam, et al. 2019; Ronneberger et al. 2015) and so-called temporal convolutional networks (Lea et al. 2016). U-Time adopts basic concepts from U-Net for 1D time-series segmentation by mapping a whole sequence to a dense segmentation in a single forward pass.

Let $\mathbf{x} \in \mathbb{R}^{\tau S \times C}$ be a physiological signal with C channels sampled at rate S for τ seconds. Let e be the frequency at which we want to segment $\mathbf{x}$, that is, the goal is to map $\mathbf{x}$ to $\lfloor \tau \cdot e \rfloor$ labels, where each label is based on $i = S/e$ sampled points. In sleep staging, 30 second intervals are typically considered (i.e., $e = 1/30$ Hz). The input x to U-Time are T fixed-length connected *segments* of the signal, each of length i. U-Time predicts the T labels at once. Specifically, the model $f(x; \theta) : \mathbb{R}^{T \times i \times C} \rightarrow \mathbb{R}^{T \times K}$ with parameters θ maps x to class confidence scores for predicting K classes for all T segments. That is, the model processes 1D signals of length $t = Ti$ in each channel.

The segmentation frequency e is variable. For instance, a U-Time model trained to segment with $e = 1/30$ Hz may output sleep stages at a higher frequency at inference time. In fact, the extreme case of $e = S$, in which every individual time-point of x gets assigned a stage, is technically possible, although difficult (or even infeasible) to evaluate (see for example Figure 8.3). U-Time, in contrast to other approaches, allows for this flexibility, because it learns an intermediate representation of the input signal where a confidence score for each of the K classes is assigned to each time point. From this dense segmentation the final predictions over longer segments of time are computed by projecting the fine-grained scores down to match the rate e at which human annotated labels are available.

The U-Time model f consists of three logical submodules: The encoder f_{enc} takes the raw physiological signal and represents it by a deep stack of feature maps, where the input is sub-sampled several times. The decoder f_{dec} learns a mapping from the feature stack back to the input signal domain that gives a dense, point-wise segmentation. A segment classifier f_{segment} uses the

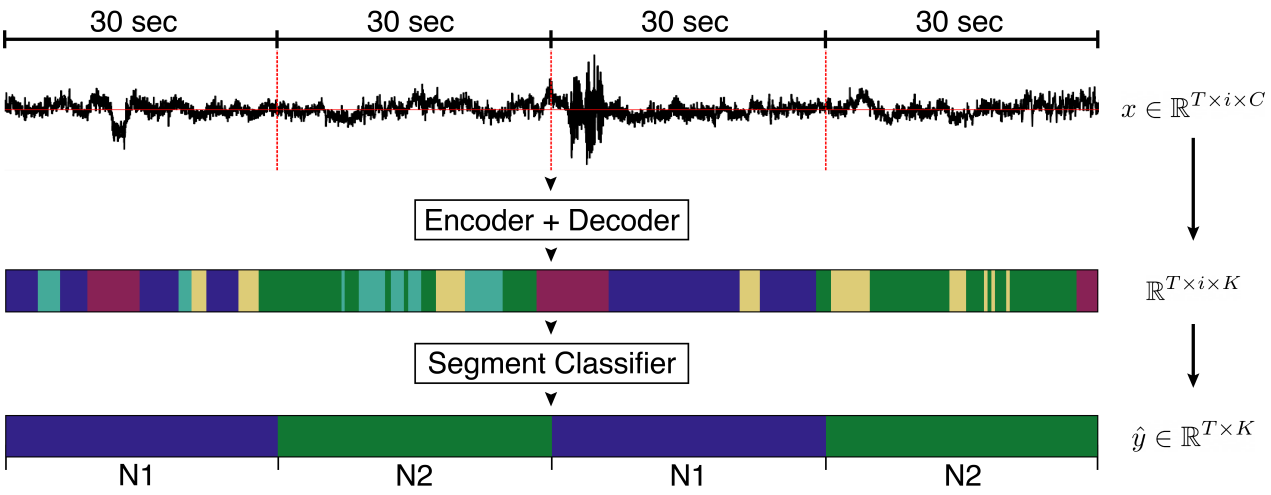


Figure 8.1: Illustrative example of how U-Time maps a potentially very long input sequence (here only $T = 4$ for visual purposes) to segmentations at a chosen temporal scale (here $e = 1/30$ Hz) by first segmenting the signal at every data-point and then aggregating these scores to form final predictions.

dense segmentation to predict the final sleep stages at a chosen temporal resolution. These steps are illustrated in Figure 8.1. An architecture overview is provided in Figure 8.2 and detailed in Supplementary Table C.2.

Encoder The encoder consists of four convolution blocks. All convolutions in the three submodules preserve the input dimensionality through zero-padding. Each block in the encoder performs two consecutive convolutions with 5-dimensional kernels dilated to width 9 (Yu et al. 2016) followed by batch normalization (Ioffe et al. 2015) and max-pooling. In the four blocks, the pooling windows are 10, 8, 6, and 4, respectively. Two additional convolutions are applied to the fully down-sampled signal. The aggressive down-sampling reduces the input dimensionality by a factor 1920 at the lowest layers. This 1) drastically reduces computational and memory requirements even for very long inputs, 2) enforces learning abstract features in the bottom layers and, 3), combined with stacked dilated convolutions, provides a large receptive field at the last convolution layer of the encoder. Specifically, the maximum theoretical receptive field of U-Time corresponds to approx. 5.5 minutes given a 100 Hz signal (see (Luo et al. 2017) for further information on theoretical and effective receptive fields).

The input x to the encoder could be an entire PSG record ($T = \lfloor \tau \cdot e \rfloor$) or a subset. As the model is based on convolution operations, the total input length t need not be static either, but could change between training and testing or even between individual mini-batches. While t is adjustable, it must be large enough so that all max-pooling operations of the encoder are defined, which in our implementation amounts to $t_{\text{min}} = 1920$ or 19.2 seconds of a 100 Hz signal. A too small t reduces performance by preventing the model from exploiting long-range temporal relations.

Decoder The decoder consists of four transposed-convolution blocks (Long et al. 2014), each performing nearest-neighbour up-sampling (Odena et al. 2016) of its input followed by convolution with kernel sizes 4, 6, 8 and 10, respectively, and batch normalization. The resulting feature maps are concatenated (along the filter dimension) with the corresponding feature maps computed by the encoder at the same scale. Two convolutional layers, both followed by batch normalization, process the concatenated feature maps in each block. Finally, a point-wise convolution with K filters (of size 1) results in K scores for each sample of the input sequence.

In combination, the encoder and decoder maps a $t \times C$ input signal to $t \times K$ confidence scores. We may interpret the decoder output as class confidence scores assigned to every sample point of the input signal, but in most applications we are not able to train the encoder-decoder network in a supervised setting as labels are only provided or even defined over segments of the input signal.

Segment classifier The segment classifier serves as a trainable link between the intermediate representation defined by the encoder-decoder network and the label space. It aggregates the sample-wise scores to predictions over longer periods of time. For periods of i time steps, the segment classifier performs channel-wise mean pooling with width i and stride i followed by point-wise convolution (kernel size 1). This aggregates and re-weights class confidence scores to produce scores of lower temporal resolution. In training, where we only have T labels available, the segment classifier maps the dense $t \times K$ segmentation to a $T \times K$-dimensional output.

Because the segment classifier relies on the mean activation over a segment of decoder output, learning the full function f (encoder+decoder+segment classifier) drives the encoder-decoder sub-network to output class confidence scores distributed over the segment. As the input to the segment classifier does not change in expectation if e (the segmentation frequency) is changed, this allows to output classifications on shorter temporal scales at inference time. Such scores may provide important insight into the individual sleep stage classifications by highlighting regions of uncertainty or fast transitions between stages on shorter than 30 second scales. Figure 8.3 shows an example.

8.4 Experiments and Evaluation

Our brain is in either an awake or sleeping state, where the latter is further divided into rapid-eye-movement sleep (REM) and non-REM sleep. Non-REM sleep is further divided into multiple states. In his pioneering work, (Kales et al. 1968) originally described four non-REM stages, S1, S2, S3 and S4. However, the American Academy of Sleep Medicine (AASM) provides a newer characterization (Iber et al. 2007), which most importantly changes the non-REM naming convention to N1, N2, and N3, grouping the original stages S3 and S4 into a single stage N3. We use this 5-class system and

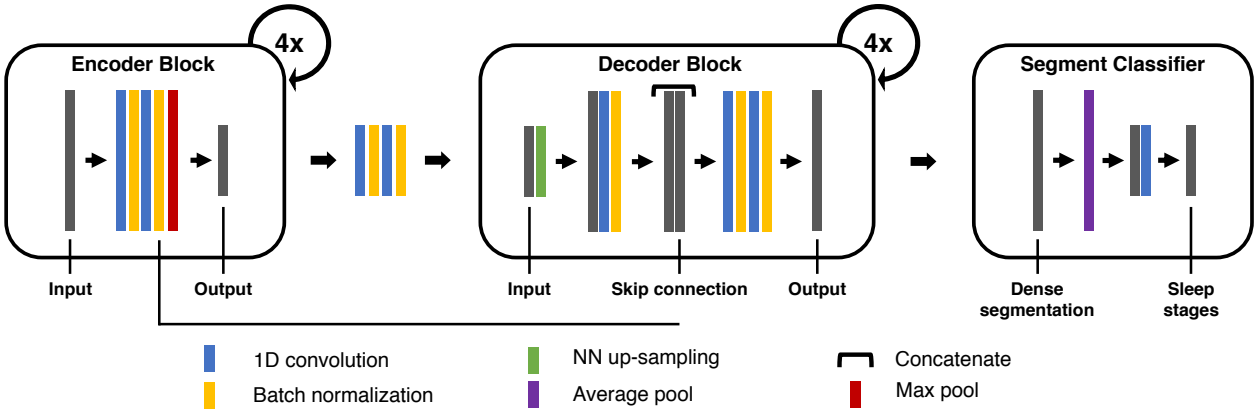


Figure 8.2: Structural overview of the U-Time architecture. Please refer to Supplementary Figure C.2 for an extended, larger version.

refer to Table C.1 in the Supplementary Material for an overview of primary features describing each of the AASM sleep stages.

We evaluated U-Time for sleep-stage segmentation of raw EEG data. Specifically, U-Time was trained to output a segmentation of an EEG signal into $K = 5$ sleep stages according to the AASM, where each segment lasts 30 seconds ($e = 1/30$ Hz). We fixed $T = 35$ in our experiments. That is, for a $S = 100$ Hz signal we got an input of $t = 105000$ samples spanning 17.5 minutes.

Our experiments were designed to gauge the performance of U-Time across several, significantly different sleep study cohorts when no task-specific modifications are made to the architecture or hyperparameters between each. In the following, we describe the data pre-processing, optimization, and evaluation in detail, followed by a description of the datasets considered in our experiments.

Preprocessing All EEG signals were re-sampled at $S = 100$ Hz using polyphase filtering with automatically derived FIR filters. Across the datasets, sleep stages were scored by at least one human expert at temporal resolution $e = 1/30$ Hz. When stages were scored according to the (Kales et al. 1968) manual, we merged sleep stages S3 and S4 into a single N3 stage to comply with the AASM standard. We discarded the rare and typically boundary-located sleep stages such as 'movement' and 'non-scored' and their corresponding PSG signals, producing the identical label set {W, N1, N2, N3, R} for all the datasets. EEG signals were individually scaled for each record to median 0 and inter quartile range (IQR) 1.

Some records display extreme values typically near the start or end of the PSG studies when electrodes are placed or the subject is entering or leaving the bed. To stabilize the pre-processing scaling as well as learned batch normalization, all 30 second segments that included one or more values higher than 20 times the global IQR of that record were set to zero. Note that this only applied if the segment was scored by the human observer (almost always classified 'wake' as these typically occur outside the 'in-bed' region), as they would otherwise be discarded. We set the values

to zero instead of discarding them to maintain temporal consistency between neighboring segments.

Optimization U-Time was optimized using a fixed set of hyperparameters for all datasets. We used the Adam optimizer (Kingma et al. 2015) with learning rate $\eta = 5 \cdot 10^{-6}$ minimizing the generalized dice cost function with uniform class weights (Crum et al. 2006; Sudre et al. 2017), $\mathcal{L}(y, \hat{y}) = 1 - \frac{2}{K} \frac{\sum_k^K \sum_n^N y_{kn} \hat{y}_{kn}}{\sum_k^K \sum_n^N y_{kn} + \hat{y}_{kn}}$. This cost function is useful in sleep staging, because the classes may be highly imbalanced. To further counter class imbalance we selected batches of size $B = 12$ on-the-fly during training according to the following scheme: 1) we uniformly sample a class from the label set {W, N1, N2, N3, R}, 2) we select a random sleep period corresponding to the chosen class from a random PSG record in the dataset, 3) we shift the chosen sleep segment to a random position within the $T = 35$ width window of sleep segments. This scheme does not fully balance the batches, as the 34 remaining segments of the input window are still subject to class imbalance.

Training of U-Time was stopped after 150 consecutive epochs of no validation loss improvement (see also Cross-validation below). We defined one epoch as $\lceil L/T/B \rceil$ gradient steps, where L is the total number of sleep segments in the dataset, T is the number of fixed-length connected segments input to the model and B is the batch size. Note that we found applying regularization unnecessary when optimizing U-Time as overfitting was negligible even on the smallest of datasets considered here (see Sleep Staging Datasets 8.4 below).

Model specification and hyperparameter selection The encoder and decoder parts of the U-Time architecture are 1D variants of the 2D U-Net type model that we have found to perform excellent across medical image segmentation problems (described in T. Koch et al. 2019; Perslev, Dam, et al. 2019). However, U-Time uses larger max-pooling windows and dilated convolution kernels. These changes were introduced in order to increase the theoretical receptive field of U-Time and were made based on our physiological understand of sleep staging rather than hyperparameter tuning. The only choice we made based on data was the loss function, where we compared dice loss and cross entropy using 5-fold cross-validation on the Sleep-EDF-39 dataset (see below). We did not modify the architecture or any hyperparameters (e.g., learning rates) after observing results on any of the remaining datasets. Our minimal hyperparameter search minimizes the risk of unintentional method-level overfitting.

U-Time as applied here has a total of ≈ 1.2 million trainable parameters. Note that this is at least one order of magnitude lower than typical CNN-LSTM architectures such as DeepSleepNet (Supratak et al. 2017). We refer to Table C.2 and Figure C.2 in the Supplementary Material for a detailed model specification as well as to Table C.3 in the Supplementary Material for a detailed list of hyperparameters.

Cross-validation We evaluated U-Time on 7 sleep EEG datasets (see below) with no task-specific architectural modifications. For a fair comparison with published results, we adopted the evaluation setting that was most frequent in the literature for each dataset. In particular, we adopted the number of cross-validation (CV) splits, which are given in the results Table 8.2 below. All reported CV scores result from single, non-repeated CV experiments.

It is important to stress that CV was always performed on a per-subject basis. The entire EEG record (or multiple records, if one subject was recorded multiple times) were considered a single entity in the CV split process.[1] On all datasets except SVUH-UCD, $\lceil 5\% \rceil$ of the training records of each split were used for validation to implement early-stopping based on the validation F1 score (Dice 1945; Sørensen 1948). For SVUH-UCD, a fixed number of training epochs (800) was used in all splits, because the dataset is too small to provide a representative validation set.

Evaluation & metrics In Table 8.2 we report the per-class F1/dice scores computed over raw confusion matrices summed across all records and splits. This procedure was chosen to be comparable to the relevant literature. The table summarizes our results and published results for which the evaluation strategy was described clearly. Specifically, we only compare to studies in which CV has been performed on a subject-level and not segment level. In addition, we only compare to studies that either report F1 scores directly or provide other metrics or confusion matrices from which we could derive the F1 score. We only compare to EEG based methods.

LSTM comparison We re-implemented the successful DeepSleepNet CNN-LSTM model (Supratak et al. 2017) for two purposes. First, we tried to push the performance of this model to the level of U-Time on the Sleep-EDF-39 and DCSM datasets (see below) through a series of hyperparameter experiments summarized in Table C.13 & Table C.14 in the Supplementary Material. Second, we used DeepSleepNet to establish a unified, state-of-the-art baseline. Because the DeepSleepNet system as introduced in (Supratak et al. 2017) was trained for a fixed number of epochs without early stopping, we argue that direct application of the original implementation to new data would favour our U-Time model. Therefore, we re-implemented DeepSleepNet and plugged it into our U-Time training pipeline. This ensures that the models use the same early stopping mechanisms, class-balancing sampling schemes, and TensorFlow implementations. We employed pre- and finetune training of the CNN and CNN-LSTM subnetworks, respectively, as in (Supratak et al. 2017). We observed overfitting using the original settings, which we mitigated by reducing the default pre-training learning rate by a factor 10. For Sleep-EDF-39 and DCSM, DeepSleepNet was manually tuned in an attempt

[1] Not doing so leads to data from the same subject being in both training and test sets and, accordingly, to overoptimistic results. This effect is very pronounced. Therefore, we do not discuss published results where training and test set were not split on a per-subject basis.

to reach maximum performance (see Supplementary Material). We did not evaluate DeepSleepNet on SVUH-UCD because of the small dataset size.

Implementation U-Time is publicly available at `https://github.com/perslev/U-Time`. The software includes a command-line-interface for initializing, training and evaluating models through CV experiments automatically distributed and controlled over multiple GPUs. The code is based on TensorFlow (Abadi et al. 2015). We ran all experiments on a NVIDIA DGX-1 GPU cluster using 1 GPU for each CV split experiment. However, U-Time can be trained on a conventional 8-12 GB memory GPU. Because U-Time can score a full PSG in a single forward-pass, segmenting 10+ hours of signal takes only seconds on a laptop CPU.

Sleep Staging Datasets We evaluated U-Time on several public and non-public datasets covering many real-life sleep-staging scenarios. The PSG records considered in our experiments have been collected over multiple decades at multiple sites using various instruments and recording protocols to study sleep in both healthy and diseased individuals. We briefly describe each dataset and refer to the original papers for details. Please refer to Table 8.1 for an overview and a list of used EEG channels.

Sleep-EDF A public PhysioNet database (Goldberger et al. 2000; B. Kemp et al. 2000) often used for benchmarking automatic sleep stage classification algorithms. As of 2019, the sleep-cassette subset of the database consists of 153 whole-night polysomnographic sleep recordings of healthy Caucasians age 25-101 taking no sleep-related medication. We utilze both the full Sleep-EDF database (referred to as *Sleep-EDF-153*) as well as a subset of 39 samples (referred to as *Sleep-EDF-39*) that correspond to an earlier version of the Sleep-EDF database that has been extensively studied in the literature. Note that for these two datasets specifically, we only considered the PSGs starting from 30 minutes before to 30 minutes after the first and last non-wake sleep stage as determined by the ground truth labels in order to stay comparable with literature such as (Supratak et al. 2017).

Physionet 2018 The objective of the 2018 Physionet challenge (Ghassemi et al. 2018; Goldberger et al. 2000) was to detect arousal during sleep from PSG data contributed by the Massachusetts General Hospital's Computational Clinical Neurophysiology Laboratory. Sleep stages were also provided for the training set. We evaluated U-Time on splits of the 994 subjects in the training set.

DCSM A non-public database provided by Danish Center for Sleep Medicine (DCSM), Rigshospitalet, Glostrup, Denmark comprising 255 whole-night PSG recordings of patients visiting the center for diagnosis of non-specific sleep related disorders. Subjects vary in demographic characteristics, diagnostic background and sleep/non-sleep related medication usage.

ISRUC Sub-group 1 of this public database (Khalighi et al. 2016) comprises all-night PSG

Table 8.1: Datasets overview. The Scoring column reports the annotation protocol (R&K = Rechtschaffen and Kales, AASM = American Academy of Sleep Medicine), Sample Rate lists the original rate (in Hz), and Size gives the number of subjects included in our study after exclusions.

Dataset	Size	Sample Rate	Channel	Scoring	Disorders
S-EDF-39	39	100	Fpz-Cz	R&K	None
S-EDF-153	153	100	Fpz-Cz	R&K	None
Physio-2018	994	200	C3-A2	AASM	Non-specific sleep disorders
DCSM	255	256	C3-A2	AASM	Non-specific sleep disorders
ISRUC	99	200	C3-A2	AASM	Non-specific sleep disorders
CAP	101	100-512	C4-A1/C3-A2	R&K	7 types of sleep disorders
SVUH-UCD	25	128	C3-A2	R&K	Sleep apnea, primary snoring

recordings of 100 adult, sleep disordered individuals, some of which were under the effect of sleep medication. Recordings were independently scored by two human experts allowing performance comparison between the algorithmic solution and human expert raters. We excluded subject 40 due to a missing channel.

CAP A public database (Terzano et al. 2002) storing 108 PSG recordings of 16 healthy subjects and 92 pathological patients diagnosed with one of bruxism, insomnia, narcolepsy, nocturnal frontal lobe epilepsy, periodic leg movements, REM behavior disorder, or sleep-disordered breathing. We excluded subjects brux1, nfle6, nfle25, nfle27, nfle33, n12 and n16 due to missing C4-A1 and C3-A2 channels or due to inconsistent meta-data information.

SVUH-UCD The St. Vincent's University Hospital / University College Dublin Sleep Apnea Database (Goldberger et al. 2000) contains 25 full overnight PSG records of randomly selected individuals under diagnosis for either obstructive sleep apnea, central sleep apnea or primary snoring.

8.5 Results

We applied U-Time with fixed architecture and hyperparameters to 7 PSG datasets. Table 8.2 lists the class-wise F1 scores computed globally (i.e., on the summed confusion matrices over all records) for U-Time applied to a single EEG channel (see Table 8.1), our re-implemented DeepSleepNet (CNN-LSTM) baseline and alternative models from literature. Table C.12 in the Supplementary material further reports a small number of preliminary multi-channel U-Time experiments, which we discuss below. Table C.5 to Table C.11 in the Supplementary Material display raw confusion matrices corresponding to the scores of Table 8.2. In Table C.4 in the Supplementary Material, we report the mean, standard deviation, minimum and maximum per-class F1 scores computed across individual EEG records, which may be more relevant from a practical perspective.

Even without task-specific modifications, U-Time reached high performance scores for large and small datasets (such as Physionet-18 and Sleep-EDF-39), healthy and diseased populations (such as

Table 8.2: U-Time results across 7 datasets. U-Time and our CNN-LSTM baseline process single-channel EEG data. Referenced models process single- or multi-channel EEG data. References: [1] Supratak et al. (2017), [2] Vilamala et al. (2017), [3] Phan, Fernando Andreotti, Cooray, Chén, et al. (2018), [4] Tsinalis et al. (2016), [5] F. Andreotti et al. (2018).

		Eval		Global F1 scores					
Dataset	Model	Records	CV	W	N1	N2	N3	REM	mean
S-EDF-39	*U-Time*	39	20	0.87	0.52	0.86	0.84	0.84	0.79
	CNN-LSTM[1]	39	20	0.85	0.47	0.86	0.85	0.82	0.77
	VGGNet[2]	39	20	0.81	0.47	0.85	0.83	0.82	0.76
	CNN[3]	39	20	0.77	0.41	0.87	0.86	0.82	0.75
	Autoenc.[4]	39	20	0.72	0.47	0.85	0.84	0.81	0.74
S-EDF-153	*U-Time*	153	10	0.92	0.51	0.84	0.75	0.80	0.76
	CNN-LSTM	153	10	0.91	0.47	0.81	0.69	0.79	0.73
Physio-18	*U-Time*	994	5	0.83	0.59	0.83	0.79	0.84	0.77
	CNN-LSTM	994	5	0.82	0.58	0.83	0.78	0.85	0.77
DCSM	*U-Time*	255	5	0.97	0.49	0.84	0.83	0.82	0.79
	CNN-LSTM	255	5	0.96	0.39	0.82	0.80	0.82	0.76
ISRUC	*U-Time*	99	10	0.87	0.55	0.79	0.87	0.78	0.77
	CNN-LSTM	99	10	0.84	0.46	0.70	0.83	0.72	0.71
	Human obs.	99	-	0.92	0.54	0.80	0.85	0.90	0.80
CAP	*U-Time*	101	5	0.78	0.29	0.76	0.80	0.76	0.68
	CNN[5]	104	5	0.77	0.35	0.76	0.78	0.76	0.68
	CNN-LSTM	101	5	0.77	0.28	0.69	0.77	0.75	0.65
SVUH-UCD	*U-Time*	25	25	0.75	0.51	0.79	0.86	0.73	0.73

Sleep-EDF-153 and DCSM), and across different EEG channels, sample rates, accusation protocols and sites etc. On all datasets, U-Time performed, to our knowledge, at least as well as any automated method from the literature that allows for a fair comparison – even if the method was tailored towards the individual dataset. In all cases, U-Time performed similar or better than the CNN-LSTM baseline.

We attempted to push the performance of the CNN-LSTM architecture of our re-implemented DeepSleepNet (Supratak et al. 2017) to the performance of U-Time on both the Sleep-EDF-39 and DCSM datasets. These hyperparameter experiments are given in Table C.13 and Table C.14 in the Supplementary Material. However, across 13 different architectural changes to the DeepSleepNet model, we did not observe any improvement over the published baseline version on the Sleep-EDF-39 dataset, indicating that the model architecture is already highly optimized for the particular study cohort. We found that relatively modest changes to the DeepSleepNet architecture can lead to large changes in performance, especially for the N1 and REM sleep stages. On the DCSM dataset, a smaller version of the DeepSleepNet (smaller CNN filters, specifically) improved performance slightly over the DeepSleepNet baseline.

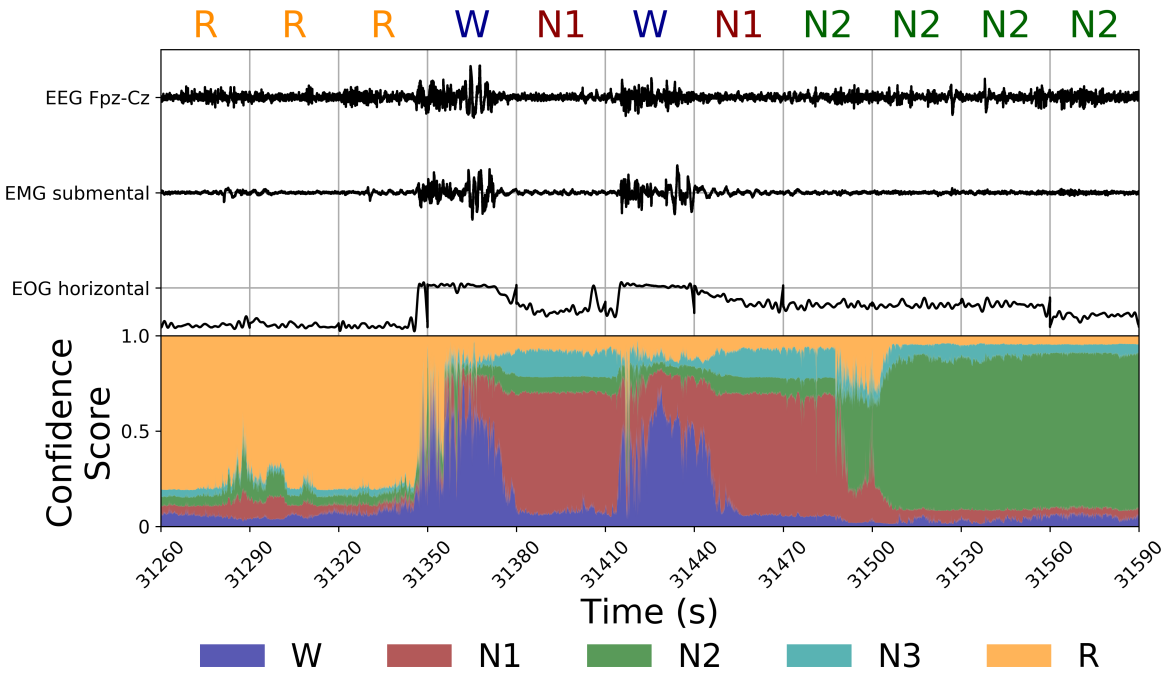


Figure 8.3: Visualization of the class confidence scores of U-Time trained on $C = 3$ input channels on the Sleep-EDF-153 dataset when the segmentation frequency e is set to match the input signal frequency. Here, U-Time outputs 100 sleep stage scores per second. The top, colored letters give the ground truth labels for each 30 second segment. The height of the colored bars in the bottom frame gives the softmax (probability-like) scores for each sleep stage at each point in time.

8.6 Discussion and Conclusions

U-Time is a novel approach to time-series segmentation that leverages the power of fully convolutional encoder-decoder structures. It first implicitly segments the input sequence at every time point and then applies an aggregation function to produce the desired output.

We developed U-Time for sleep staging, and this study evaluated it on seven different sleep PSG datasets. For all tasks, we used the same U-Time network architecture and hyperparameter settings. This does not only rule out overfitting by parameter or structure tweaking, but also shows that U-Time is robust enough to be used by non-experts – which is of key importance for clinical practice. In all cases, the model reached or surpassed state-of-the-art models from the literature as well as our CNN-LSTM baseline. In our experience, CNN-LSTM models require careful optimization, which indicates that they may not generalize well to other cohorts. This is supported by the observed drop in CNN-LSTM baseline performance when transferred to, for example, the ISRUC dataset. We further found that the CNN-LSTM baseline shows large F1 score variations, in particular for sleep stage N1, for small changes of the architecture (see Table C.13 in the Supplementary Material). In contrast, U-Time reached state-of-the-art performance across the datasets without being tuned for each task. Our results show that U-Time can learn sleep staging based on various input channels

across both healthy and diseased subjects. We attribute the general robustness of U-Time to its fully convolutional, feed-forward only architecture.

Readers not familiar with sleep staging should be aware that even human experts from the same clinical site may disagree when segmenting a PSG.[2] While human performance varies between datasets, the mean F1 overlap between typical expert annotators is at or slightly above 0.8 (Stephansen et al. 2018). This is also the case on the ISRUC dataset as seen in Table 8.2. U-Time performs at the level of the human experts on the three non-REM sleep stages of the ISRUC dataset, while inferior on the REM sleep stage and slightly below on the wake stage. However, human annotators have the advantage of being able to inspect several channels including the EOG (eye movement), which often provides important information in separating wake and REM sleep stages. This is because the EEG activity in wake and REM stages is similar, while – as the name suggests – characteristic eye movements are indicative of REM sleep (see Table C.1 in the Supplementary Material). In this study we chose to use only a single EEG channel to compare to other single-channel studies in literature. It is highly likely that U-Time for sleep staging would benefit from receiving multiple input channels. This is supported by our preliminary multi-channel results reported in Supplementary Table C.12. On ISRUC and other datasets, the inclusion of an EOG channel improved classification of the REM sleep stage.

We observed the lowest U-Time performance on the CAP dataset, although on par with the model of (F. Andreotti et al. 2018), which requires multiple input channels. The CAP dataset is difficult because it contains recordings from patients suffering from seven different sleep related disorders, each of which are represented by only few subjects, and because of the need for learning both the C4-A1 and C3-A2 channels simultaneously.

Besides its accuracy, robustness, and flexibility, U-Time has a couple of other advantageous properties. Being fully feed-forward, it is fast in practice as computations may be distributed efficiently on GPUs. The input window T can be dynamically adjusted, making it possible to score an entire PSG record in a single forward pass and to obtain full-night sleep stage classifications almost instantaneously in clinical practice. Because of its special architecture, U-Time can output sleep stages at a higher temporal resolution than provided by the training labels. This may be of importance in a clinical setting for explaining the system's predictions as well as in sleep research, where sleep stage dynamics on shorter time scales are of great interest (H. Koch, Poul Jennum, et al. 2019). Figure 8.3 shows an example.

While U-Time was developed for sleep staging, we expect its basic design to be readily applicable to other time series segmentation tasks as well. Based on our results, we conclude that fully

[2]This is true in particular for the N1 sleep stage, which is difficult to detect due to its transitional nature and non-strict separation from the awake and deep sleep stages.

convolutional, feed-forward architectures such as U-Time are a promising alternative to recurrent architectures for times series segmentation, reaching similar or higher performance scores while being much more robust with respect to the choice of hyperparameters.

Chapter 9

Paper D

U-Sleep: Resilient high-frequency sleep staging

Authors

Mathias Perslev[a], Sune Darkner[a], Lykke Kempfner[b], Miki Nikolic[b], Poul Jørgen Jennum[b] and Christian Igel[a].

[a]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

[b]Danish Center for Sleep Medicine, Rigshospitalet, Glostrup, Denmark

Published

npj Digital Medicine 4, 2021. DOI: https://doi.org/10.1038/s41746-021-00440-5

9.1 Abstract

Sleep disorders affect a large portion of the global population and are strong predictors of morbidity and all-cause mortality. Sleep staging segments a period of sleep into a sequence of phases providing the basis for most clinical decisions in sleep medicine. Manual sleep staging is difficult and time-consuming as experts must evaluate hours of polysomnography (PSG) recordings with electroencephalography (EEG) and electrooculography (EOG) data for each patient. Here we present U-Sleep, a publicly available, ready-to-use deep-learning-based system for automated sleep staging (sleep.ai.ku.dk). U-Sleep is a fully convolutional neural network which was trained and evaluated on PSG recordings from 15,660 participants of 16 clinical studies. It provides accurate segmentations across a wide range of patient cohorts and PSG protocols not considered when building the system. U-Sleep works for arbitrary combinations of typical EEG and EOG channels, and its special deep learning architecture can label sleep stages at shorter intervals than the typical 30 second periods used during training. We show that these labels can provide additional diagnostic information and lead to new ways of analyzing sleep. U-Sleep performs on par with state-of-the-art automatic sleep staging systems on multiple clinical datasets, even if the other systems were built specifically for the particular data. A comparison with consensus-scores from a previously unseen clinic shows that U-Sleep performs as accurately as the best of the human experts. U-Sleep can support the sleep staging workflow of medical experts, which decreases healthcare costs, and can provide highly accurate segmentations when human expertise is lacking.

9.2 Introduction

Sleep disorders affect a large portion of the global population and impose significant welfare costs (Devore et al. 2016; P. Jennum et al. 2009; Silva et al. 2016; Tobaldini et al. 2018; Wittchen et al. 2011). Abnormal sleeping patterns and associated sleep disorders are strong predictors of morbidity and all-cause mortality (Chattu et al. 2018; Garbarino et al. 2016). Anomalous sleep-wake changes occur for instance in psychiatric conditions (e.g., schizophrenia, depression Baandrup et al. 2018), neurodegenerative diseases (e.g., dementia, rapid eye movement sleep behavior disorder, and Parkinson's Disease Baandrup et al. 2018; H. Koch, Poul Jennum, et al. 2019; Olesen et al. 2018), and genuine sleep disorders (e.g., narcolepsy (Stephansen et al. 2018), insomnia (Miller et al. 2016), sleep apnea H. Koch, Schneider, et al. 2017) as well as during epileptic seizures and prior to stroke (Ponsaing et al. 2017). Timely and accurate diagnosis of sleep disorders relies on the difficult and time-consuming process of sleep staging based on polysomnography (PSG) data. A PSG collects a set of non-invasive long-term recordings of physiological measures of multiple brain and body functions using modalities such as electroencephalography (EEG), electrooculography (EOG), and electromyography (EMG). These signals are divided into intervals, typically of 30 seconds, which are mapped to different sleep stages such as awake, light sleep, intermediate sleep, deep sleep, and rapid eye movement (REM) sleep (Iber et al. 2007; Kales et al. 1968) (see Supplementary Figure C.1 and Supplementary Table C.1 for a brief overview of PSG and sleep stage characteristics). This sleep staging forms the basis for subsequent analyses.

Sleep staging requires multiple hours of manual annotations from expert clinicians for each subject incurring significant costs and leading to bottlenecks in both diagnosis and large-scale clinical studies. The manual annotations suffer from high intra- and interscorer variability which reduces the diagnostic precision (Stephansen et al. 2018; Warby et al. 2014). Algorithmic sleep staging aims at automating this process. Recent work shows that such systems can be highly accurate and robust and may play an important role in developing novel biomarkers for sleep disorders and other (e.g., neurodegenerative) diseases (Anderer et al. 2010; Klosh et al. 2001; Schenck et al. 2014; Stephansen et al. 2018; Warby et al. 2014). Deep learning (LeCun, Bengio, et al. 2015) is becoming increasingly popular for the analysis of physiological time-series in general (Faust, Hagiwara, et al. 2018) and has already been successfully applied to sleep staging (Faust, Razaghi, et al. 2019; Robert et al. 1998; Ronzhina et al. 2012). While several high-performance deep-learning-based sleep staging systems have been proposed recently (Biswal, Sun, et al. 2018; Chambon et al. 2018; Dong et al. 2018; Guillot et al. 2019; Kuo et al. 2020; Mousavi et al. 2019; Phan, Fernando Andreotti, Cooray, O. Y. Chen, et al. 2019; Phan, Chén, P. Koch, Lu, et al. 2020; Phan, Chén, P. Koch, Mertins, et al. 2020; Sun et al. 2017; Supratak et al. 2017), these have not yet been widely adopted in clinical practice because

it is not clear if the reported results can be generalized. Current state-of-the-art systems are tuned, trained and evaluated on one or a very small number of clinical cohorts, and it remains questionable if similar results can be achieved in a different clinical setting for different patient populations. Most systems are designed to operate on PSG data from a specific hardware & pre-processing pipeline including a specific set of EEG/EOG/EMG channels, sampling rate, etc. to maximize performance. Consequently, most existing sleep staging systems – including deep learning systems trained on several datasets (F. Andreotti et al. 2018; Phan, Chén, P. Koch, Lu, et al. 2020) – require re-training at each clinical site, which imposes a significant technical barrier.

A robust, easy-to-use sleep staging model directly applicable across clinical populations and PSG protocols with (at least) expert-level performance would both free significant resources across sleep clinics and enable developing countries with advanced sleep diagnostics. Such a system may also serve as a global, standardized reference for sleep staging which could spark scientific discussions and reduce inter-clinical and inter-operator variability.

This study describes U-Sleep, our contribution towards these goals. U-Sleep is a publicly available, ready-to-use deep neural network for resilient sleep staging inspired by the popular U-Net (Brandt et al. 2020; Falk et al. 2019; Ronneberger et al. 2015) architecture for image segmentation. The neural network was trained and evaluated on the – to the best of our knowledge – largest and most diverse set of PSG records for sleep staging ever collected, spanning 16 independent clinical studies providing 23 datasets, geographically dispersed clinical sites, multiple decades, a large array of demographics, and patient groups. Eight datasets were not considered during model development and training, they were only used for realistic verification of U-Sleep. Two datasets are consensus-scored and allowed us to compare U-Sleep's performance to that of five clinical experts on both healthy subjects and sleep-disordered patients. U-Sleep requires only a single EEG and a single EOG channel with arbitrary standard electrode placement as input, makes no assumptions about the acquisition hardware (including sampling rate) or pre-processing pipeline, and outputs a whole night's sleep stages in seconds on a laptop CPU. U-Sleep also has a unique in-built ability to output sleep stage labels at temporal frequencies up to the signal sampling rate (Perslev, Jensen, et al. 2019). We show that such high-frequency representation of sleep carries diagnostic information in separating obstructive sleep apnea (OSA) patients from a population of healthy control subjects.

Figure 9.1 provides an overview of the U-Sleep prediction pipeline. Figure 9.2 illustrates the model architecture. U-Sleep is freely available at `https://sleep.ai.ku.dk`.

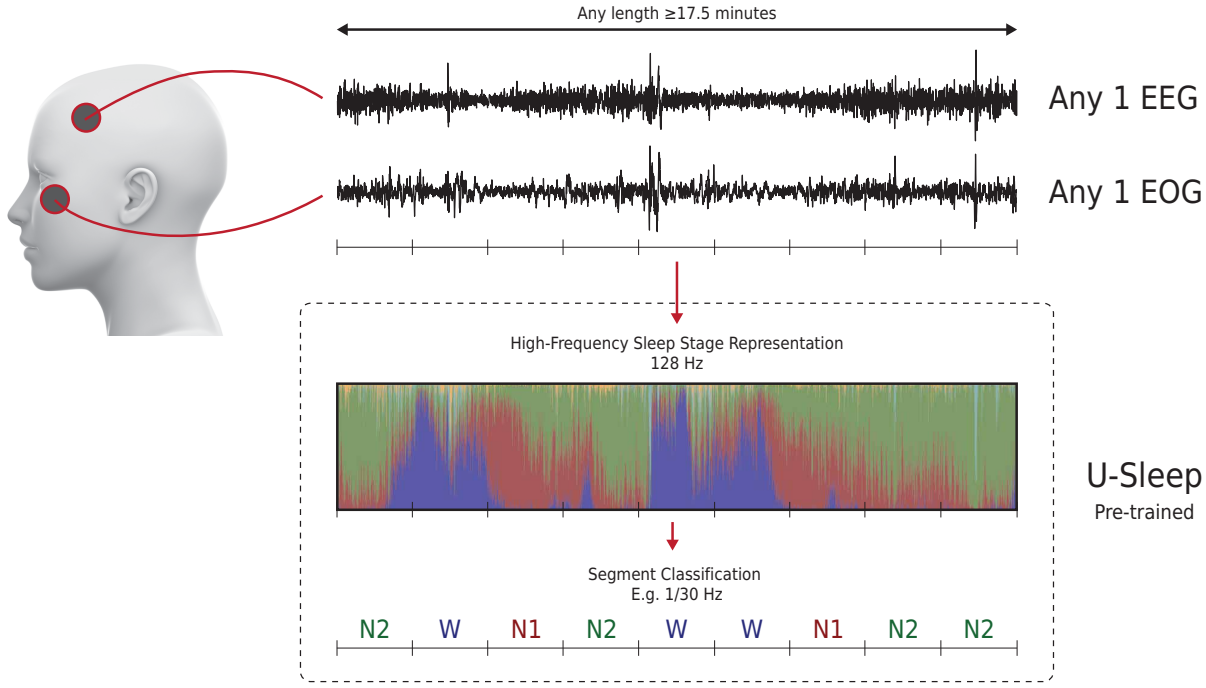


Figure 9.1: The U-Sleep prediction pipeline. U-Sleep is a ready-to-use deep neural network for sleep staging. First, it maps each provided EEG and EOG channel pair (shown in the top) to an intermediate, high-frequency sleep stage representation (shown in the middle). The intermediate representation is visualized by the colored bars indicating the level of confidence U-Sleep has that the subject is in one of the 5 sleep stages at a given time (Blue: Wake, Red: N1, Green: N2, Cyan: N3, Yellow: REM). From the intermediate representation, U-Sleep aggregates confidence scores over periods of time (for instance segments of 30 seconds) to output final sleep staging scores. U-Sleep makes no assumptions about the PSG protocol including acquisition hardware, electrode positions, filtering, and sampling rate. Internally, signals are re-sampled at 128 Hz. U-Sleep may output sleep stage labels up to this frequency.

9.3 Methods

9.3.1 Fully Convolutional Neural Network for Time Series Segmentation

The U-Sleep model is a deep neural network which maps an EEG and an EOG signal to a high-frequency sleep stage representation and then aggregates this intermediate representation to a sequence of sleep stages each spanning a fixed-length time interval (e.g., 30 seconds). This process is illustrated in Figure 9.1. U-Sleep accepts input signals obtained with any common electrode placement (i.e., any EEG and EOG channel), hardware and software filtering, and sampling rate (internally re-sampled to 128 Hz). Up to computer memory constraints, U-Sleep processes inputs of arbitrary lengths. However, inputs shorter than 17.5 minutes may reduce performance by restricting the model from observing long-range dependencies in the data. and predicts sleep stages for the whole sequence in a single forward pass. This makes it possible for U-Sleep to process a whole night's PSG data in seconds on commodity hardware and even in less than a second if a graphics processing unit (GPU) is used.

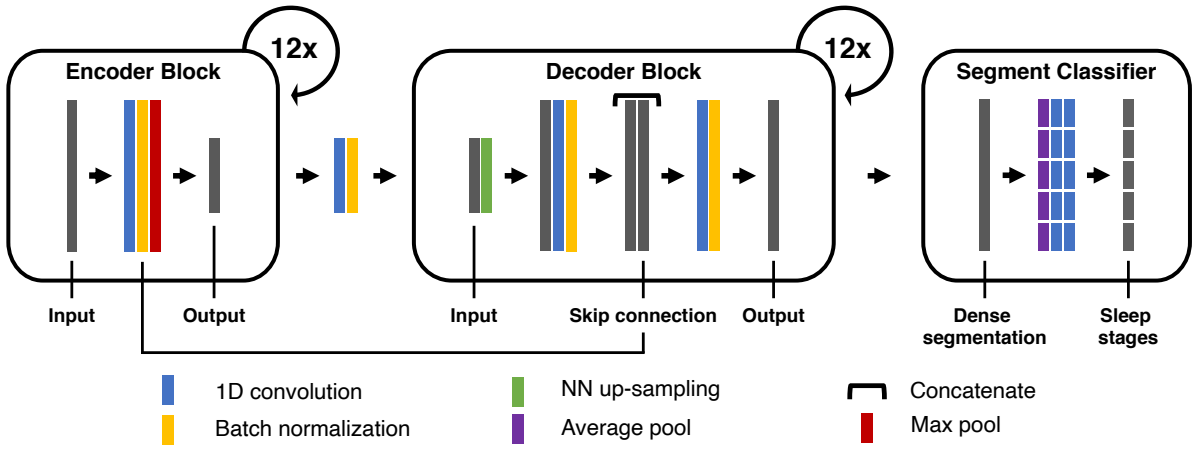


Figure 9.2: Model architecture. U-Sleep is a special fully convolutional neural network architecture designed for physiological time-series segmentation tasks such as sleep staging. It consists of an encoder (left) which encodes the input signals into dense feature representations, a decoder (middle) which projects the learned features into the input space to generate a dense sleep stage representation as shown in Figure 9.1, and finally a specially designed segment classifier (right) which generates sleep stages at a chosen temporal resolution. Please see the Method section and Supplementary Table D.1 for details on the U-Sleep model architecture.

In contrast to other automated sleep staging systems, U-Sleep is a purely feed-forward, fully convolutional neural network. Fully convolutional networks have been incredibly successful in computational vision and especially in medical image analysis. They mark the state of the art in image segmentation, with the *U-Net* arguably being the most popular architecture so far (Falk et al. 2019; Ronneberger et al. 2015). We successfully applied U-Nets for various medical segmentation tasks, and found that one fixed architecture and set of hyperparameters can give excellent results across very different tasks (A. Desai et al. 2020; T. Koch et al. 2019; Perslev, Dam, et al. 2019). Recently, we adapted our version of *U-Net* for image analysis (Brandt et al. 2020; A. Desai et al. 2020; T. Koch et al. 2019; Perslev, Dam, et al. 2019) to the segmentation of one-dimensional physiological time series data. We extended the architecture with an additional block of fully convolutional layers for aggregating classifications (Perslev, Jensen, et al. 2019). The new architecture termed *U-Time* was applied to sleep staging.

In accordance with our results on images, we found that we could use the same network architecture and training process to learn a variety of sleep staging tasks outperforming state-of-the-art models such as DeepSleepNet (Supratak et al. 2017). Our fully convolutional network was easier to train (e.g., less dependent on hyperparameter settings) compared to more complex models for sleep staging relying on recurrent neural network architectures (Perslev, Jensen, et al. 2019). Another decisive feature of U-Time is that it provides a classification of the input signals for each time point as an intermediate representation, although the data used for training and evaluating the model were segmented at a much lower temporal resolution. The U-Sleep architecture proposed in this study

supersedes U-Time; the main differences between the systems are described below.

9.3.2 Automated Sleep Staging

Sleep staging refers to the process of partitioning a PSG record into a sequence of sleep stages. Human annotators typically consider segments of 30 seconds and assign a single sleep stage to each segment. We denote a PSG record by $\mathbf{X} \in \mathbb{R}^{\tau S \times C}$, where τ is a number of seconds sampled, S is the sampling rate and C is the number of channels recorded. The output of the sleep staging process is a sequence of $\lfloor \tau \cdot e \rfloor$ labels, where e is the frequency at which we want to assign sleep stages, with $e = 1/30$ Hz being the typical value for human annotators. Thus, each sleep stage spans $i = S/e$ sampled points in time across C channels.

Given a fixed integer $i > 0$, U-Sleep defines a deterministic function $f(\mathbf{X}'; \theta) : \mathbb{R}^{T \cdot i \times C} \rightarrow \mathbb{R}^{T \times K}$ for any integer $T > 0$, where θ is a set of parameters learned from data, $\mathbf{X}'$ is a (section of a) PSG record, T is a number of fixed-length segments with i sampled points each, C the number of PSG channels and K is the number of sleep stages. During training, $\mathbf{X}'$ is typically a submatrix of a longer PSG $\mathbf{X}$ with $\mathbf{X}' = \mathbf{X}[\{t, \dots, t + i \cdot T\}, \{1, \dots, C\}]$ for some time point t. That is, U-Sleep takes a temporal section of a PSG and outputs a sequence of labels corresponding to fixed-length, contiguous segments of time (in principle, different output labels of U-Sleep could span different lengths of time, but we assume the typical case of fixed-length segments). The input $\mathbf{X}'$ can be any length (augmented or cut to a multiple of i; ideally $T \cdot i \geq 4096$, because there are 12 pooling operations down-sampling the signal by a factor of 2 each). For instance, when we trained U-Sleep, $\mathbf{X}'$ spanned 17.5 minutes of a PSG signal. When using U-Sleep to predict sleep stages in new data, the whole PSG is input to U-Sleep (i.e., $\mathbf{X}' = \mathbf{X}$), which computes the whole hypnogram at once.

The provided U-Sleep system requires at least two input channels ($C = 2$), one EEG and one EOG channel, respectively, sampled or re-sampled to 128 Hz. It assumes $K = 5$ different stages {Wake, N1, N2, N3, REM}.

9.3.3 Machine Learning Model

U-Sleep is a fully convolutional deep neural network refining its predecessor U-Time (Perslev, Jensen, et al. 2019), which we recently devised for time-series segmentation problems such as sleep staging (the differences between U-Sleep and U-Time are described below). In the following, we outline the U-Sleep architecture. We refer to Figure 9.2 for a schematic overview and to Supplementary Table D.1 for additional details on the configuration of the individual layers.

U-Sleep consists of three sub-modules: 1) An *encoder* module first extracts a deep stack of abstract feature maps from the input signals. Each extracted feature map has a lower temporal

resolution compared to its input. 2) A *decoder* module then performs an up-scaling of the compact feature maps to match the temporal resolution of the input signals. The output of the decoder may be seen as a complex representation of sleep stages at a frequency matching the input signal. 3) A specially designed *segment classifier* module aggregates the intermediate, high-frequency output of the decoder into segments and predicts the sleep stages for these segments. For each segment, a confidence score is predicted for every possible sleep stage, which is interpreted as a probabilistic prediction by applying the softmax-function. Next, we describe the individual modules in more detail.

Encoder The encoder module comprises 12 *encoder blocks*. Each encoder block consists of one convolutional layer (kernel size 9, no kernel dilation, stride 1 Goodfellow et al. 2016), one layer of Exponential Linear Unit (ELU) (Clevert et al. 2016) activation functions, batch normalization (Ioffe et al. 2015) and max-pooling (kernel size 2, stride 2). The number of learned filters c_l in the l-th convolutional layer is $\sqrt{2}$ times larger compared to the previous layer, starting with $c_1 = 5$, that is, for $l \in \{1, ..., 11\}$ we have $c_{l+1} = \lfloor c_l \sqrt{2} \rfloor$ (this corresponds to a doubling of the degrees of freedom from one block to the next, which is less than in U-Net implementations).

Decoder The decoder module consists of 12 *decoder blocks*. Each decoder block performs nearest neighbour up-sampling of the input with kernel size 2 (i.e., it doubles the length of the feature maps along the temporal axis) and applies convolution (kernel size 2, stride 1), ELU activation functions and batch normalization. The up-scaled input is then combined with the output of the batch-norm operation (i.e., before max-pooling) of the corresponding encoder block (in terms of temporal resolutions, e.g., the first decoder block matches the last encoder block). Then a convolution, non-linearity, and batch-normalization are applied to the stacked feature maps. Opposite to the encoder, the decoder scales *down* the number of learned filters by a factor of $\sqrt{2}$ in each consecutive block.

The output of the final decoder has the same temporal resolution as the input signal. Thus, when concatenated, the encoder and decoder modules map an input signal in $\mathbb{R}^{T \cdot i \times C}$ to an output in $\mathbb{R}^{T \cdot i \times K}$, where $K = 5$ is the number of sleep stages. This output can be regarded as an intermediate representation of sleep stages at high (128 Hz) frequency.

Segment Classifier The segment classifier module maps the intermediate, high-frequency representation to the sleep stage prediction at the desired frequency. It aggregates scores over longer segments of time. For a given window of length i it first applies a per-channel mean-pooling operation with kernel width i and stride i. Two point-wise convolution operations (kernel width 1, stride 1) are then applied, the first using ELU activation functions. This allows to learn a non-linear weighted combination of the mean scores over the interval. Finally, the softmax function is used to transform

the scores into probabilistic predictions. Thus, the output of the segment classifier is a $T \times K$ right stochastic matrix, where T is a number of segments and $K = 5$ is the number of sleep stages. During training, we have one sleep stage label available for each segment of length i, and we train the whole encoder + decoder + segment classifier network end-to-end as described in the Optimization section below.

9.3.4 Model Specification and Hyperparameter Selection

The deep neural network architecture of U-Sleep is well-structured and simple in comparison to many others deep networks proposed for sleeep staging. Still, the U-Sleep has many hyperparameters (e.g., the depth, the number of filters and their sizes for each block, etc.) which could be optimized to tune its performance on any specific set of data. However, we deliberately did not systematically tune the hyperparameters of U-Sleep, but employed a minimal hyperparameter selection strategy based on empirical evidence gathered from the U-Time (Perslev, Jensen, et al. 2019), our experience from using fully convolutional neural networks for image segmentation (Brandt et al. 2020; T. Koch et al. 2019; Perslev, Dam, et al. 2019), and our physiological understanding of sleep staging. We avoided automated hyperparameter search to limit unintentional method-level overfitting and problems due to adaptive data analysis.

We adopted large parts of the U-Sleep model architecture (Supplementary Table D.1) and hyperparameters (Supplementary Table D.2) from its predecessor U-Time (Perslev, Jensen, et al. 2019), which was shown to be able to learn sleep staging across a range of datasets (individually) without requiring dataset-specific hyperparameter tuning. Still, we changed important aspects of the system. Because U-Sleep solves a significantly more difficult learning task requiring generalization across clinical cohorts and input channel combinations we increased the capacity of the network. The increased dataset size allowed us to fit a more complex model. In addition, we improved the system based on lessons learnt from U-Time. U-Sleep has a larger number of trainable parameters ($\approx 3.1 \cdot 10^6$ compared to U-Time's $\approx 1.1 \cdot 10^6$) and is significantly deeper, consisting of 12 encoder- and decoder blocks instead of U-Time's four. U-Sleep also down-samples the input signal and subsequent feature maps much more slowly by using max-pooling kernels of width 2 in all encoder blocks instead of U-Time much more aggressive max-pooling kernels of widths in $\{10, 8, 4, 2\}$. U-Sleep implements the more complex ELU non-linearity following all convolution operations instead of U-Time's Rectified Linear Units (ReLUs). Finally, whereas U-Time only linearly combined the mean-pooled activations in the final segment classifier layer, U-Sleep applies two convolution operations allowing for a non-linear weighted combination.

All changes served to increase the capacity of U-Sleep (i.e., its ability to approximate a more complex target function). Using a less aggressive max-pool down-sampling strategy reduces the

information loss in the early layers. While U-Time benefited from early, aggressive down-sampling to reach computational and statistical efficiency, we argued that U-Sleep might need to capture more complex, hardly conceived patterns in the input signals which are robustly observed across datasets and channel combinations but may be lost if the input is sub-sampled too aggressively. The increased depth of U-Sleep also considerably expanded its theoretical receptive field (Luo et al. 2017) (the maximum length of input signal that may effect each convolution computation in a given layer) from U-Time's ≈ 5.5 minutes to ≈ 9.6 minutes in the last convolutional layer of the encoding sub-network. We numerically estimated the output sleep stages to be sensitive to changes in the input space 6.75 minutes backward and forward in time (i.e., each sleep stage prediction is informed by data from a window of up to 13.5 minutes of 128 Hz signal).

While U-Sleep has more layers compared to U-Time, the individual encoder- and decoder blocks are less complex, because they apply only a single convolution operation to their inputs instead of two, and the number of learned filters scale only by a factor of $\sqrt{2}$ with depth instead of 2 (see Supplementary Table D.2).

Finally, we trained U-Sleep differently from U-Time to accommodate learning across many different datasets, and also apply augmentations as described in the Optimization and Augmentation sections below. The more common and simpler cross-entropy cost function was optimized instead of the generalized dice loss (Crum et al. 2006; Sudre et al. 2017) used for U-Time.

All reported results in this study are from the first and only trained instance of the U-Sleep model. That is, the design choices described above were not revised based on the performance of the system, making the reported evaluation metrics highly reliable.

9.3.5 Pre-processing

All EEG and EOG signals are resampled to 128 Hz using polyphase filtering. We scale the range of EEG and EOG signals on a global, per-subject and per-channel basis so that the whole EEG signal recorded from a single channel has a median of 0 and inter quartile range (IQR) of 1 (i.e. an outlier robust scaling). We then clip any value which has an absolute deviation from the median of more than 20 times the IQR of that channel. Finally, during training we strip from the beginning and end any EEG or EOG signal which is outside the range of the scored hypnogram.

The current U-Sleep system considers sleep stages following the AASM standard: { W, N1, N2, N3, REM } (Iber et al. 2007). If data was originally scored by a human expert following the Kales and Rechtschaffen (Kales et al. 1968) manual, we merged stages S3 and S4 into a single N3. U-Sleep does not attempt to score stages such as 'MOVEMENT' or 'UNKNOWN'. Whenever such a label occurred during training, we masked the loss contribution from that segment. This ensures that the model observes the segment in question, but its prediction does not influence the computation

of the gradients for updating the model. We did not remove such segments entirely, as we want a model that can deal with such potentially noisy regions when scoring neighbouring segments after deployment.

9.3.6 Augmentation

Data augmentation refers to modifying the input data during training to improve generalization. We applied transformations to a random subset of the sampled batch elements, replacing variable lengths of segments within EEG and EOG channels or even entire channels with Gaussian noise. Specifically, for each sample in a batch, with probability 0.1, a fraction of the signals in that sample was replaced with noise from $N(\mu = \hat{\mu}, \sigma^2 = 0.01)$, where $\hat{\mu}$ is the empirically measured mean of the sample's signals. The fraction was sampled log-uniformly from $[0.001, ..., 0.3]$. With probability 0.1 at most 1 channel was entirely replaced by noise. These augmentations were applied to force the model to consider both channels and complex distant relations in the signal.

9.3.7 Input Channel Majority Voting

When applying U-Sleep to new PSG data we utilize its ability to accept input data from arbitrary EEG and EOG electrode positions by predicting the full hypnogram for each combination of 1 EEG and 1 EOG channel possible for the given PSG. The resulting predictions are then combined to one final hypnogram. For each segment, the softmax scores (values ranging from 0 to 1 indicating the model's confidence in each sleep stage) of all predictions are summed up and the sleep stage with the highest accumulated score is the final prediction for the segment.

The hypnogram based on an ensemble of predictions is likely to be more accurate than the individual hypnograms, as multiple predictions may smooth out errors if those are uncorrelated across channels (Masegosa et al. 2020; Perslev, Dam, et al. 2019) and provide additional evidence to difficult, borderline cases.

9.3.8 Evaluation

U-Sleep outputs sleep stages in $\{\mathrm{W}, \mathrm{N1}, \mathrm{N2}, \mathrm{N3}, \mathrm{REM}\}$ as described above. When evaluating U-Sleep we scored the full PSG, but did not consider the predicted class on a segment with a label different from the five sleep stages (e.g., a segment labelled 'MOVEMENT' or, for whatever reason, not scored by a human expert at all). We predicted sleep stages using all combinations of available EEG and EOG channels for each PSG. Unless otherwise specified, we used majority voting fusing these predictions when evaluating U-Sleep. We refer to the supplementary material for channel-wise results.

We evaluated U-Sleep using the F1/Sørensen-Dice metric (Dice 1945; Sørensen 1948), which is computed for each sleep stage c separately. The F1 score is defined as $\mathrm{F}^c_{\beta=1} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$, where TP, FP and FN are the number of true positives, false positives and false negatives for a given class c. The F1 score is used, because it emphasises both recall and precision. We computed the F1 score for all 5 classes from (non-normalized) confusion matrices and report them separately or combined by calculating the unweighted mean. Note that unweighted F1 scores typically reduce the absolute scores due to lower performance on less abundant classes such as sleep stage N1.

Table 9.2 gives an overview over the results, reporting only F1 scores computed for a given class across all subjects of a testing set, which results in a single number without error bars. In Table 9.3 we consider F1 scores computed for each subject individually and report the mean and standard deviation, which may better reflect performance in a clinical setting.

Each PSG record in the datasets `DOD-H` and `DOD-O` was scored by 5 human experts. This allows us to compute consensus-scored hypnograms that may be regarded as ground truth data and then evaluate the performance of U-Sleep in relation to this ground truth as well as in comparison to individual human experts. We used the code provided with the `DOD` publication (see `https://github.com/Dreem-Organization/dreem-learning-evaluation`) for evaluating the consensus scores (Guillot et al. 2019), except that we did not balance the F1 scores measured for each class by the abundance of that class (we report unweighted mean F1 scores for consistency reasons). When comparing a human annotator to the consensus, the consensus hypnograms are generated from the $N-1$ remaining expert scores. In accordance to the literature, U-Sleep and other automated methods reported in Table 9.3 were evaluated against consensus hypnograms based on the $N-1$ most reliable annotators (Guillot et al. 2019).

9.3.9 High-Frequency Sleep Staging Experiments

U-Sleep has the ability to make predictions at higher temporal resolutions compared to the the labels used during training. As an intermediate representation, U-Sleep computes a confidence score for each possible sleep stage at each sampled time point (i.e., at 128 Hz in the current system). An example of this is shown in Figure 9.1. Sleep stages are inherently defined based on patterns observed over (longer) time periods. Thus, the question is whether the high-resolution outputs are informative of actual physiological sleeping patterns or only add more noise.

During training, our model considers the mean of the confidence scores over a 30-seconds segment, shuffling the scores within a segment would not change the learning and the prediction. Still, it is likely that the intermediate scores will reflect the true sleep stage at a time point, because only in that way the system can be independent of the – to a large extend arbitrary – positioning of the windows defining the segments.

One way to assess the usefulness of the scores is by linking them to a clinical diagnosis. We considered the datasets `DOD-H` and `DOD-O` (see Table 9.1 and the Supplementary Material Datasets section) with 25 healthy subjects and 55 obstructive sleep apnea (OSA) patients, respectively. As OSA patients suffered from abrupt awakenings and rapid transitions from deep sleep into lighter sleep stages, we expected a classifier to be able to separate the two populations with better-than-random performance given simple features describing the number of such transitions per time. For each subject in `DOD-H` and `DOD-O`, we predicted sleep stages at frequencies in $\{2, 4, 8, 16, 32, 64, 128, 256, 512, 768, 1280\}$ predictions/min. We used all available combinations of EEG and EOG channels (16 for DOD-O and 24 for DOD-H) and computed the majority vote for each segment. For each subject we considered the 2 predictions/minute output for determining the onset and end of sleep (indicated by first and last sleep stage). For all frequencies, only the sleep stages within this time-frame were considered.

For each segment of 1.5 hours of sleep we counted the number of occurrences of sleep stage transition triplets. A triplet is a sequence $(s_1, s_2, s_3) \in \{\mathrm{W}, \mathrm{N1}, \mathrm{N2}, \mathrm{N3}, \mathrm{REM}\}^3$. We considered only triplets for which $s_1 \neq s_2$ and $s_2 \neq s_3$. This leaves 80 different triplets in which a fast transition to stage s_2 occur (e.g., $(\mathrm{N3}, \mathrm{W}, \mathrm{N1})$) ignoring more typical triplets such as $(\mathrm{N2}, \mathrm{N2}, \mathrm{N2})$.

We fit a random forest classifier (Breiman 2001) (using the `sklearn` implementation, Pedregosa et al. 2011) to the triplet frequencies (occurrences per time). We fit the classifier to 79 out of the 80 subjects and predicted whether the last subject suffers from OSA or not, repeating the process for all subjects (leave-one-out cross validation). We repeated the whole experiment 50 times for each frequency bin with a small randomization in the hyperparameters of the random forest classifier. The latter is done to increase our confidence that any observed correlation is not related to a very specific set of hyperparameters. Specifically, in each repetition of the experiment we trained a random forests with 200 trees with respect to the Gini impurity measure and class weights $w_c = n/(k \cdot n_c)$, where w_c is the weight associated with class c, n is the total number of samples, n_c the number of samples of class c, and $k = 2$ is the number of classes ('balanced' mode in `sklearn` notation). A random value was chosen for the following hyperparameters: `maximum_tree_depth` $\in \{2, \ldots, 7\}$, `min_samples_leaf` $\in \{2, ..., 7\}$, `min_samples_split` $\in \{2, ..., 7\}$ and `max_features` $\in \{\mathrm{sqrt}, \mathrm{log2}\}$. We refer to `https://scikit-learn.org/stable/modules/ensemble.html#forest` for a detailed description of those parameters.

We determined the overall OSA classification for each subject by the majority vote over the predictions of the model across all segments of 1.5 hours of sleep, ties were broken at random.

9.4 Results

9.4.1 Datasets and Model Training

We trained and evaluated U-Sleep on 19,924 PSG records collected from 15,660 participants of 16 independent clinical studies (21 datasets). A brief overview of each dataset along with key demographic statistics are displayed in Table 9.1, details can be found in Supplementary Note: Datasets. All datasets are publicly available, some require an approval. The datasets can be split into two groups. First, there are 13 datasets that were partly used to train the U-Sleep model. In combination they span ≈ 19.4 years of annotated signals. Each dataset was split into training (at least 75 %), validation (up to 10 %, at most 50 subjects) and testing (up to 15 %, at most 100 subjects) subsets on a per-subject or per-family basis. All records from subjects in the training sets were used to train the U-Sleep model. Records in the validation sets where used to monitor the performance of U-Sleep throughout training. Records in the testing subsets were used for evaluation.

In the second group are 8 datasets that were used for evaluation only, that is, no data from these sources were used in the model building process (neither for training nor hyperparameter selection). These datasets allowed an unbiased performance evaluation of U-Sleep when applied (unaltered) to new, clinical cohorts. Among others, we measured the performance of U-Sleep against human experts by considering held-out consensus-scored datasets produced by clinical experts. The performance of U-Sleep was compared to that of the individual experts evaluated against their consensus scores.

The combined training dataset spans a significant fraction of the expected clinical population including large numbers of healthy individuals, patients with sleep and non-sleep related disorders, men and women, as well as different age-, BMI- and ethnic groups. The datasets were collected across geographically diverse locations (although mainly from the US), across decades, and on a variety of hardware using different sampling rates, hardware filters and more. The datasets were scored by sleep experts with different backgrounds.

Our goal is to perform accurate sleep staging across all cohorts simultaneously. In contrast to most studies, we deliberately exposed the machine learning system to highly variable data and labels, in order to learn a final model which generalizes well and is useful in clinical practice where data may vary unexpectedly and with time. U-Sleep was trained on randomly selected batches sampled across the datasets as described in the Methods section. For each sample in a batch, U-Sleep was exposed to a randomly selected EEG and EOG channel combination picked from all possible combinations for the given PSG. No information was given to the model about the data sources. This challenging setup forced U-Sleep to become invariant to electrode placements. We designed U-Sleep to require only a single EEG and a single EOG channel, where the electrode placement does

not matter as long as it is a standard position, to maximize its applicability and ease-of-use. We omitted other modalities such as EMG, which carry important information about sleep disturbances and disorders (e.g., Parkinson's Disease and REM sleep behavior disorder), but are not necessary for the delineation of sleep stages. Adding EMG has the potential to further improve the performance of U-Sleep. However, EMG signals especially help to distinguish between being awake and REM sleep, two stages that our predecessor system U-Time already separates very well. In preliminary experiments, adding EMG did not improve the performance of U-Time (see supplementary Table S.12 in the study of Perslev, Jensen, et al. 2019). Using only the two most common modalities makes our model widely applicable, in particular in scenarios without advanced sleep monitoring setups, and allowed us to combine many datasets for training, some of which did not, for example, contain EMG recordings.

9.4.2 Performance Overview

U-Sleep was able to learn sleep staging across all training datasets simultaneously. Supplementary Figure D.2 shows both the overall loss and mean F1 score computed across validation subsets for each individual training dataset. The U-Sleep performance increased at similar rates for all datasets.

We used the trained U-Sleep to predict the full hypnogram of all PSG records in the test subsets of all datasets using all available combinations of EEG and EOG channels. Given the large number of results, we focus on the mean and stage-wise F1/Dice metrics computed across subjects for each dataset as described in the Methods section. The per-channel evaluations are shown for each dataset in Supplementary Tables D.3–D.24. Table 9.2 lists the F1 scores using majority vote, that is, the hypnograms were generated using predictions from all available EEG-EOG channel combinations within each record. Majority voting, as can be seen from the channel-wise results in the Supplementary Material, always performed at least as good as the average over all possible channel combinations. For 19 out of the 21 datasets, the majority voting performed at least as good as the best individual channel (see Supplementary Material).

Across 21 datasets, U-Sleep performed sleep staging with mean F1 $\pm$ STD (in parenthesis shown when weighted by number of test records) of 0.90 ± 0.04 (0.91 ± 0.03) for stage Wake, 0.53 ± 0.07 (0.53 ± 0.07) for stage N1, 0.85 ± 0.04 (0.86 ± 0.03) for N2, 0.76 ± 0.07 (0.77 ± 0.08) for stage N3, and 0.90 ± 0.02 (0.90 ± 0.02) for stage REM. Considering the mean computed across stages for each dataset, the global F1 performance can be summarized as 0.79 ± 0.03 (0.79 ± 0.03) ranging from a minimum 0.73 (`SVUH`) to maximum 0.85 (`CCSHS` and `CHAT`). The standard deviation over F1 scores obtained using each available channel combination was for most datasets below 0.02 (mean 0.01), with datasets `MASS-C1` and `ABC` being the only exceptions with standard deviations of 0.03.

Examples of hypnograms as computed by U-Sleep using channel majority voting are visualized

and compared to human expert annotations for all 21 testing datasets in in Supplementary Figures D.3–D.23. Specifically, we display the predicted hypnogram with the single highest F1-score, the single lowest F1-score and the one nearest to the median F1 score observed for the dataset. Thus, the figures visualize the span in U-Sleep performance from worst- to best-case scenario.

9.4.3 Consensus Results: Comparing to Human Experts

In Table 9.3 we report the performance of U-Sleep on the consensus-scored datasets `DOD-H` (9.3a) and `DOD-O` (9.3b) compared to the performance of 5 individual clinical experts from which the consensus scores were generated. The distributions of scores are shown for U-Sleep and the 5 experts in Figure 9.3.

Across the 25 healthy subjects of `DOD-H`, U-Sleep matched the best performing human expert with a mean F1 score of 0.79 ± 0.07 and human expert scores ranging from a minimum 0.72 ± 0.11 (Expert 4) to a maximum 0.79 ± 0.07 (Expert 3). There is no significant difference between the performance of U-Sleep and the best human expert (Expert 3) at confidence level $\alpha = 0.05$ ($W = 150.0$, $p = 0.737$, two-sided Wilcoxon test). U-Sleep scored higher mean F1 than all humans on stages Wake and N1, similar to the best individual expert (Expert 3) on stage REM and worse than all human experts on stage N3. U-Sleep performed on average on par with the best two models *SimpleNet* and *DeepSleepNet* from the six models evaluated in the publication presenting the data (Guillot et al. 2019) (3 best shown here), which were trained on consensus-scored labels from the same data distribution.

Across the 55 OSA patients of `DOD-O`, U-Sleep had the highest mean performance of 0.76 ± 0.10 among the set of human experts and itself, with human performances ranging from a minimum 0.69 ± 0.12 (Expert 1) to a maximum 0.74 ± 0.11 (Expert 5). There is no significant difference between the performance of U-Sleep and the best human expert (Expert 5, $W = 555.0$, $p = 0.072$, two-sided Wilcoxon test). U-Sleep scored higher mean F1 than all humans on stages N1, N3 and REM, and slightly below Expert 5 on stages Wake and N2. U-Sleep performed as well or better than the reference models, which were trained on consensus labels.

9.4.4 Evaluation of High Frequency Sleep Stages

U-Sleep can output sleep stages at a higher frequency than that of the labels used during trained. We trained with a label frequency of 1/30 Hz – the most typical so called *page size* in sleep staging – but can provide sleep stage predictions at frequencies up to 128 Hz (input records may be sampled at a higher frequency, but will be re-sampled before analysis). Figure 9.1 visualizes these high-frequency scores. We argue that these scores can capture sleeping patterns on shorter time scales (Perslev,

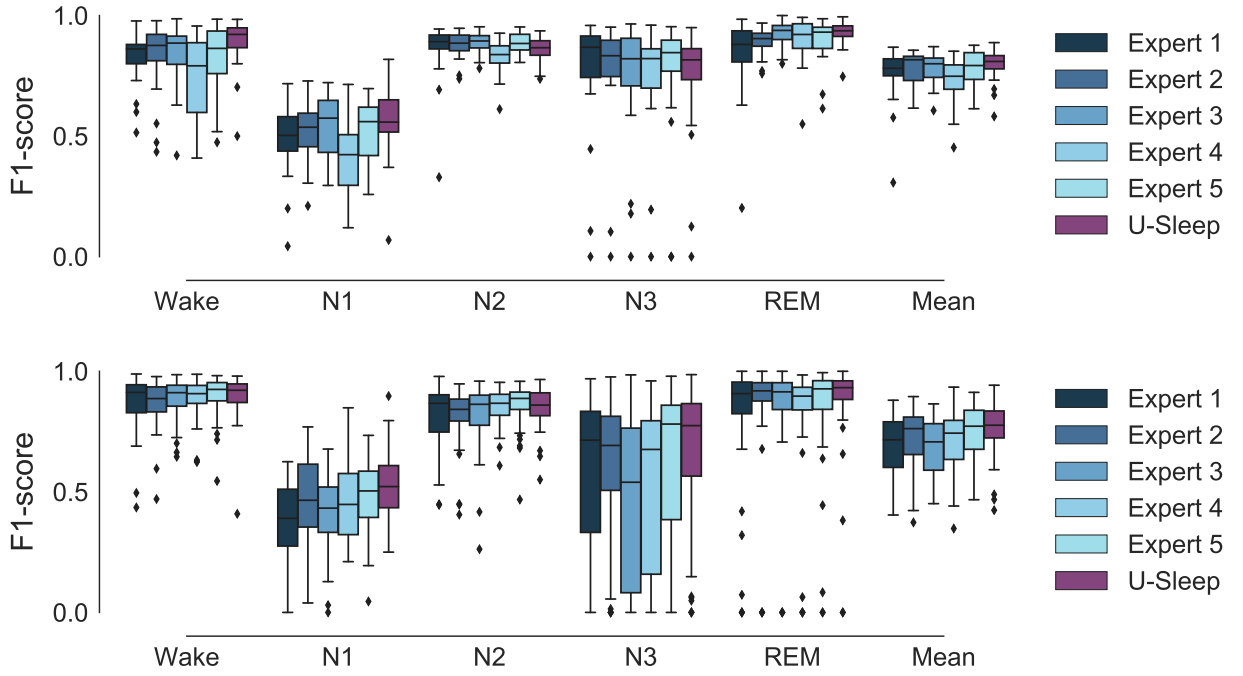


Figure 9.3: Boxplots illustrating the distributions of F1 scores from 5 human experts and U-Sleep on healthy controls and OSA patients. Panel (a) shows results from dataset `DOD-H` on 25 healthy subjects. Panel (b) shows results from dataset `DOD-O` on 55 patients suffering from OSA. Sleep stages produced by U-Sleep and the 5 individual experts were compared to consensus-scored hypnograms. Please refer to the Methods section for further details. Mean F1 scores averaged across stages are shown along with F1 scores for the 5 individual sleep/wake stages. The performance of U-Sleep is shown in red colors (right most boxplot in each group). The performance of each human expert is shown in shades of blue (4 left most boxplots in each group). Note that some records were scored by both human experts and U-Sleep with very low F1 scores (0 in some cases) on individual classes. This especially concerns stage N3 in dataset `DOD-O` and most often happens for rare classes. For instance, a patient severely affected by OSA rarely enters the N3 deep sleep stage, and the resulting low number of observed N3 stages makes even a few errors result in a large deviation in the F1 score. Each box-plot shows the median (middle vertical line), first and third quartiles (lower and upper box limits) and whiskers that extend to 1.5 times the IQR added or subtracted the third and first quartiles, respectively. Data outside of this range is marked as outliers indicated by diamond shaped points.

Jensen, et al. 2019). To show this, we performed a simple, but carefully designed study to investigate if there is predictive information in the high frequency scores. We describe the experimental details in the Methods section. We considered the datasets `DOD-H` and `DOD-O`. Our experiment evaluated the hypothesis that the healthy subjects and OSA patients are easier to discriminate by a classifier when extracting features from high-resolution sleep stage scores. We considered the output by U-Sleep at different frequencies and computed the occurrences of sleep-stage triplet transitions of the form (s_1, s_2, s_3), where $s_x \in \{\text{Wake}, \text{N1}, \text{N2}, \text{N3}, \text{REM}\}$ and $s_1 \neq s_2$ and $s_2 \neq s_3$. The extracted triplet frequency features are time-invariant. We get the same number of features independent of the frequency at which we computed them. We fitted Random Forrest (Breiman 2001) classifiers to

separate the healthy and OSA patients using features extracted at different frequencies.

Figure 9.4 shows the result of the experiment. We evaluated the classification performance on sleep stages generated by U-Sleep at 14 different frequencies approximately uniformly distributed on a $\log_2$ scale from 2 stages/minute to 7680 stages/minute (128 Hz). The mean F1 classification performance increased from an initial low value of 0.60 (at 2 stages/minute frequency) up to a maximum of 0.94 (at 1280 stages/minute), indicating that the task of separating healthy and OSA patients was much easier using high-frequency scores, and, consequently, that such stages are indeed clinically informative.

9.5 Discussion

U-Sleep has simultaneously learned sleep staging for a wide range of clinical cohorts without requiring adaptation to different cohorts. It can deal with large variations in patient demographics and PSG protocols, and only requires an arbitrary single EEG and EOG channel as inputs. We evaluated U-Sleep on several datasets that it has not seen during training, and we found that its accuracy matches the performance of models that were specifically developed and/or trained on these datasets. For instance, U-Sleep matches the performance of its predecessor U-Time (for a performance comparison of U-Time with other sleep staging approaches we refer to Perslev, Jensen, et al. 2019) trained specifically on datasets `ISRUC-SG1` and `SVUH` (Perslev, Jensen, et al. 2019) with both models scoring 0.77 and 0.73 mean F1, respectively. U-Sleep also approximately matches the performance of U-Time on datasets `SEDF-SC` (U-Sleep: 0.79, U-Time: 0.76), `PHYS` (U-Sleep: 0.79, U-Time: 0.77) and `DCSM` (U-Sleep: 0.81, U-Time: 0.79). However, the scores of U-Time on these datasets span additional records, so the results cannot be compared directly. U-Sleep performs nearly as well as DeepSleepNet on `MASS-C3` (Supratak et al. 2017) (mean F1 of 0.82 for DeepSleepNet and 0.80 for U-Sleep). It is even as accurate as the best human expert of a group of five when evaluated on the datasets `DOD-H` and `DOD-O` with healthy and diseased individuals. It performs at least as well as all six automated systems evaluated in the original study presenting these data (Guillot et al. 2019). In contrast to U-Sleep, these six models were all trained on the same consensus-scored labels that define the ground truth, which gives them the advantage of learning from higher quality labels as well as a matching label distributions at training and test time.

In contrast to other automated systems, U-Sleep is trained to work with any standard EEG and EOG channels it receives as input. The measured F1 scores do vary between individual channels (as seen in Supplementary Tables D.8–D.24), but with a low standard deviation for most datasets. Prediction by combining the available channels using majority vote almost always matches the prediction using the best individual channel. As majority scores can be easily obtained also in practice,

this result reliefs sleep researchers from testing which channel combinations work best for their specific patients and data. In accordance with our clinical experience, we did not find specific EEG and EOG channel combinations that score particularly well or badly across datasets. It is possible that the performance scores obtained using a specific channel reflect what information the human annotators focused on when annotating the signals, as individual experts may have personal preference (or training) when detecting certain sleep stage characteristics such as spindles and K-complexes in a particular set of channels. As U-Sleep is trained on randomly varying channel combinations, it is forced to learn robust features that are conceivable across EEG channels. We hypothesize that U-Sleep utilizes its ability to look minutes into both the past and future to detect more global sleeping patterns that are observable across channels, but may be difficult to conceive for humans.

Developing sleep staging systems based on deep learning is an active research area, and new findings will further improve U-Sleep. When trained on single datasets, some recent algorithms may perform better than the general U-Sleep system (Kuo et al. 2020; Phan, Chén, P. Koch, Mertins, et al. 2020). While recent work showed that specialised systems can be applied to new datasets with good performance using transfer learning techniques (F. Andreotti et al. 2018; Phan, Chén, P. Koch, Lu, et al. 2020), these methods were retrained on new data matching the target cohort, which requires technical expertise, time, specialized hardware, and labelled data from the target domain. However, no system has demonstrated the robustness of U-Sleep on the much more difficult and relevant task of resilient sleep staging across new clinical cohorts, different input channels, etc. without additional training.

The U-Sleep architecture was designed based on our experience with U-Net-type neural networks (Brandt et al. 2020; A. Desai et al. 2020; T. Koch et al. 2019; Perslev, Dam, et al. 2019), please refer to the Methods section for details. It is a limitation of our study that, because of the long training time and in order to avoid problems due to adaptive data analysis, we have not fine-tuned the U-Sleep architecture and training procedure. It is possible that small modifications could further improve the results. Also, while we have attempted to compile as many and diverse datasets as possible (e.g., with respect to demographics), all datasets used so far were collected in either Europe or North America, and represent in particular healthy subjects and OSA patients; two groups both likely to display normal EEG patterns. It remains to be systematically studied how U-Sleep performs on patients with highly abnormal brain activity patterns (e.g., following stroke or due to psychiatric diseases or neurodegenerative disorders). In addition, we have only limited patient record information available for all subjects. Accordingly, it has not been possible to fully rule out all potential (e.g., regional) biases of U-Sleep. It is our hope that more sleep data will be made available from currently underrepresented groups of subjects, training on which will reduce the risk of unintentional biases. In Supplementary Note: Demographic Bias, we report the effects of age, sex and gender,

finding increasing age to have a statistically significant (but small in magnitude) negative effect on performance, which we attribute to general decrease in health with age.

U-Sleep is an accurate, carefully evaluated and ready-to-use system for sleep staging. Therefore, we believe that the public availability of U-Sleep will benefit researchers and clinicians in sleep medicine. It can augment the workflow of expert clinicians by immediately providing sleep stage annotations of high quality when a PSG sample is inspected. While we do not advocate to disregard the invaluable expertise of the local clinical and technical staff, who will undoubtedly have a significantly better understanding and experience with patients and data from their clinic, we think that significant resources can be saved by using U-Sleep's predictions as a starting point for sleep staging. In this case, the expert only needs to spot potential errors or disagreements with the system's output instead of scoring the whole PSG manually. Furthermore, U-Sleep can provide highly accurate sleep staging when experienced experts are missing. It computes in seconds on a laptop CPU and requires no technical expertise to use, which makes it applicable for home-monitoring and sleep clinics in developing countries.

U-Sleep may facilitate large-scale, global studies of sleep with more consistent and less biased labels. While manual sleep staging follows guidelines as suggested by, for example, the AASM (Iber et al. 2007), it is a difficult process with room for interpretation, making it inconsistent and error prone (Warby et al. 2014). Different clinics may perform sleep staging slightly differently, which may introduce systematic biases when pooling data from clinical sites. While U-Sleep may make errors, these are more consistent. U-Sleep could thus be used to annotate large collections of data from across the world, facilitating the on-going and presumably important transition to large-scale sleep studies (Bragazzi et al. 2019). Individual clinics may be interested to compare their scores against those of U-Sleep, which may spark scientific debate about observed differences.

The ability of U-Sleep to output high-frequency sleep stages has the potential to significantly impact the study of sleeping patterns in health and disease, as demonstrated by our proof-of-concept experiments separating OSA patients and healthy controls. The current standard for sleep classification has developed only little since its first formulation in 1968 (Kales et al. 1968), in particular given the great progress made towards understanding sleep physiology. Sleep staging today almost always considers the brain as if it would move discretely from one stage to another over segments of exactly 30 seconds, failing to account for sleep dynamics on shorter time scales (H. Koch, Poul Jennum, et al. 2019). As we have shown, sleep stage scores at much higher frequency may serve as the basis for building future diagnostic predictive models. Such models – which may take additional input modalities such as EMG and demographic variables into account – may require significantly less training data compared to models that must learn to solve a predictive diagnostic task from raw PSG data alone, because they can utilize that U-Sleep has already digested the complex raw signals

into an informative, high-frequency representation of sleep.

Ethical Approval

The Research Ethics Committee for SCIENCE and HEALTH, University of Copenhagen, has reviewed this research project and has found it compliant with the relevant Danish and International standards and guidelines for research ethics. The `DCSM` dataset was extracted and anonymized by the Danish Center for Sleep Medicine under a general approval from the Danish Data Protection Agency to analyze historical PSG data. All other datasets were acquired from third party databases and handled according to the relevant data sharing agreements.

Data Availability

We make the DCSM dataset publicaly available at `https://sid.erda.dk/wsgi-bin/ls.py?share_id=fUH3xbOXv8`. This repository will be frozen and issued a DOI for persistent access following the review process. All other datasets are in principle also publicly available assuming the individual researcher and use-case is eligible for a given dataset as determined by the third-party dataset licence holders listed for each dataset individually in the Supplementary Material. Please refer to Table 9.1 for an overview of which datasets require approval and which are directly available.

Confusion matrices for U-Sleep predictions on all channel combinations (including majority votes) for individual subjects in all test datasets may be downloaded from `https://sid.erda.dk/wsgi-bin/ls.py?share_id=HE5nA4Xs37`. These matrices allow re-computation of F1 metrics as reported here, as well as other metrics of interest. The repository also stores hyperparameter configuration files as well as dataset preprocessing and splitting information needed to reproduce the training of U-Sleep.

Code Availability

The in-house developed codebase used for training U-Sleep is publicaly available on GitHub at `https://github.com/perslev/U-Time`. The codebase is supplied with the submitted manuscript. The software includes a command-line-interface for initializing, training and evaluating models without the need to alter the underlying codebase. The software is based on `TensorFlow` (Abadi et al. 2015). Please refer to the `README` file of the repository for guidance on installation and a step-by-step guide on how to train a U-Sleep model on a subset or all of the datasets considered here. We trained U-Sleep on a single GPU (NVIDIA Titan X) with 12 GiB of memory. Because U-Sleep can score

a full PSG in a single forward pass, segmenting 10+ hours of signal takes only seconds on a laptop CPU and is practically instantaneous if running on a GPU.

We make inference using the pre-trained U-Sleep model freely available at `https://sleep.ai.ku.dk` for non-commercial usage. Users may upload (anonymised or public domain) PSG files (European Data Format, EDF, or HDF5) to the service, choose parameters such as which channels to use and the inference frequency (e.g., 1/30 Hz or higher), and receive back the automatically scored hypnogram. The service also provides a simple interface to interactively visualize the scored hypnogram and to obtain key sleep statistics over selected periods of time. The raw sleep stages can be downloaded in several formats. We welcome community feedback on how we may improve the service with additional features.

Acknowledgements

We gratefully acknowledge support from the Independent Research Fund Denmark through the project „U-Sleep“ (project number 9131-00099B).

The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

The Apnea, Bariatric surgery, and CPAP study (ABC Study) was supported by National Institutes of Health grants R01HL106410 and K24HL127307. Philips Respironics donated the CPAP machines and supplies used in the perioperative period for patients undergoing bariatric surgery.

The Cleveland Children's Sleep and Health Study (CCSHS) was supported by grants from the National Institutes of Health (RO1HL60957, K23 HL04426, RO1 NR02707, M01 Rrmpd0380-39).

The Cleveland Family Study (CFS) was supported by grants from the National Institutes of Health (HL46380, M01 RR00080-39, T32-HL07567, RO1-46380).

The Childhood Adenotonsillectomy Trial (CHAT) was supported by the National Institutes of Health (HL083075, HL083129, UL1-RR-024134, UL1 RR024989).

The Home Positive Airway Pressure study (HomePAP) was supported by the American Sleep Medicine Foundation 38-PM-07 Grant: Portable Monitoring for the Diagnosis and Management of OSA.

The Multi-Ethnic Study of Atherosclerosis (MESA) Sleep Ancillary study was funded by NIH-NHLBI Association of Sleep Disorders with Cardiovascular Health Across Ethnic Groups (RO1 HL098433). MESA is supported by NHLBI funded contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168 and N01-HC-95169 from the National Heart, Lung, and Blood Institute, and by cooperative agreements UL1-TR-000040, UL1-TR-001079, and UL1-TR-

001420 funded by NCATS.

The National Heart, Lung, and Blood Institute provided funding for the ancillary MrOS Sleep Study, "Outcomes of Sleep Disorders in Older Men," under the following grant numbers: R01 HL071194, R01 HL070848, R01 HL070847, R01 HL070842, R01 HL070841, R01 HL070837, R01 HL070838, and R01 HL070839.

The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University).

The Study of Osteoporotic Fractures (SOF) was supported by National Institutes of Health grants (AG021918, AG026720, AG05394, AG05407, AG08415, AR35582, AR35583, AR35584, RO1 AG005407, R01 AG027576-22, 2 R01 AG005394-22A1, 2 RO1 AG027574-22A1, HL40489, T32 AG000212-14).

The ISRUC-SLEEP Dataset has been supported by the Portuguese Foundation for Science and Technology (FCT) under Ph.D. Grants SFRH/BD/81828/2011 and SFRH/BD/80735/2011 and by QRENfunded project SLEEPTIGHT, with FEDER reference CENTRO-01-0202-FEDER-011530. We would also like to acknowledge sleep experts from the Sleep Medicine Centre of the Hospital of Coimbra University (CHUC), for their invaluable support in data acquisition and visually scoring the PSG recordings of ISRUC-Sleep dataset.

Table 9.1: Datasets overview. Please refer to the Supplementary Material for additional details on each dataset. Missing values are due to study design or anonymized data. Individual statistics may be computed over a smaller number of observations than the total number of subjects due to missing data. Datasets `DOD-H` and `DOD-O` are hold-out consensus scored datasets. ABC = Apnea, Bariatric surgery, and CPAP (Bakker et al. 2018), CCSHS = Cleveland Children's Sleep and Health Study (Rosen, Larkin, et al. 2003), CFS = Cleveland Family Study (Redline, Tishler, et al. 1995), CHAT = Childhood Adenotonsillectomy Trial (C. L. Marcus et al. 2013; Redline, Amin, et al. 2011), HPAP = Home Positive Airway Pressure (Rosen, Auckley, et al. 2012), MESA = Multi-Ethnic Study of Atherosclerosis (X. Chen et al. 2015), MROS = Osteoporotic Fractures in Men (Blackwell et al. 2011; Song et al. 2015), SHHS = Sleep Heart Health Study (Quan et al. 1998), and SOF = Study of Osteoporotic Fractures (Cummings et al. 1990; Spira et al. 2008), PHYS = 2018 PhysioNet/CinC Challenge (Ghassemi et al. 2018), SEDF = Sleep-EDF (B. Kemp et al. 2000), SVUH = St. Vincent's University Hospital / University College Dublin Sleep Apnea Database (Goldberger et al. 2000), DCSM = Danish Centre for Sleep Medicine, ISRUC = ISRUC-Sleep (Khalighi et al. 2016), MASS = The Montreal Archive of Sleep Studies (O'Reilly et al. 2014), and DOD = Dreem Open Datasets (Arnal et al. 2019; Guillot et al. 2019; Thorey et al. 2019). (✓) = requires approval. *Number of distinct families. **Assuming uniform age distribution in the binned data.

Type	Dataset	Public	Records	Subjects	Length (days)	Age (years)	BMI	Sex % (F/M)
Internal - Train/Test	ABC	(✓)	132	49	46.2	48.8 ± 9.8	38.9 ± 2.9	43/57
	CCSHS	(✓)	515	515	240.1	17.7 ± 0.4	25.1 ± 5.9	50/50
	CFS	(✓)	730	730/144*	300.8	41.7 ± 20.0**	32.4 ± 9.5	55/45
	CHAT	(✓)	1638	1232	679.6	6.6 ± 1.4	19.0 ± 4.9	52/48
	DCSM	✓	255	255	201.02	–	–	–
	HPAP	(✓)	238	238	77.6	46.5 ± 11.9	37.3 ± 9.2	43/57
	MESA	(✓)	2056	2056	905.5	69.4 ± 9.1	–	54/46
	MROS	(✓)	3926	2903	1877.3	76.4 ± 5.5	27.2 ± 3.8	0/100
	PHYS	✓	994	994	309.8	55.2 ± 14.3	–	33/67
	SEDF-SC	✓	153	78	144.1	58.8 ± 22.0	–	53/47
	SEDF-ST	✓	44	22	14.8	40.2 ± 17.7	–	68/32
	SHHS	(✓)	8444	5797	3144.4	63.1 ± 11.2	28.2 ± 5.1	52/48
	SOF	(✓)	453	453	188.1	82.8 ± 3.1	27.7 ± 4.7	100/0
Hold-Out	ISRUC-SG1	✓	100	100	31.3	51.1 ± 15.9	–	44/56
	ISRUC-SG2	✓	16	8	4.9	46.9 ± 17.5	–	25/75
	ISRUC-SG3	✓	10	10	3.1	39.6 ± 9.6	–	10/90
	MASS-C1	(✓)	53	53	19.9	63.6 ± 5.3	–	36/64
	MASS-C3	(✓)	62	62	21.8	42.5 ± 18.9	–	55/45
	SVUH	✓	25	25	7.2	50.0 ± 9.4	31.6 ± 3.9	16/84
	DOD-H	✓	25	25	8.6	35.3 ± 7.5	23.8 ± 3.4	24/76
	DOD-O	✓	55	55	18.5	45.6 ± 16.5	29.6 ± 6.4	36/64

Table 9.2: Majority vote results overview. For each record in each dataset, U-Sleep generated a hynogram using all possible combinations of 1 EEG and 1 EOG channel. Results reported here are from the majority voted hypnograms across all such combinations as described in the Methods section. We refer to Supplementary Tables D.3-D.24 for per-channel results. Here we report the global F1 scores across all subjects in each dataset. The reported Mean (weighted) and STD (weighted) statistics are computed across datasets in each column weighted by the number of PSG records in each row.

Type	Dataset	Records	Wake	N1	N2	N3	REM	Mean
Internal - Train/Test	ABC	20	0.87	0.53	0.84	0.72	0.90	0.77
	CCSHS	78	0.93	0.63	0.91	0.88	0.93	0.85
	CFS	92	0.93	0.52	0.89	0.84	0.91	0.82
	CHAT	128	0.93	0.64	0.87	0.90	0.90	0.85
	DCSM	39	0.97	0.48	0.86	0.83	0.89	0.81
	HPAP	36	0.91	0.48	0.84	0.78	0.90	0.78
	MESA	100	0.92	0.59	0.87	0.65	0.90	0.79
	MROS	134	0.93	0.46	0.87	0.68	0.88	0.77
	PHYS	100	0.84	0.60	0.84	0.81	0.87	0.79
	SEDF-SC	23	0.93	0.57	0.86	0.71	0.88	0.79
	SEDF-ST	8	0.80	0.58	0.88	0.64	0.91	0.76
	SHHS	140	0.93	0.51	0.87	0.76	0.92	0.80
	SOF	68	0.93	0.45	0.86	0.77	0.92	0.78
Hold-Out	ISRUC-SG1	100	0.89	0.52	0.79	0.77	0.88	0.77
	ISRUC-SG2	16	0.85	0.49	0.78	0.83	0.86	0.76
	ISRUC-SG3	10	0.90	0.55	0.78	0.74	0.85	0.77
	MASS-C1	53	0.94	0.41	0.81	0.61	0.88	0.73
	MASS-C3	62	0.93	0.54	0.87	0.75	0.91	0.80
	SVUH	25	0.80	0.37	0.81	0.78	0.88	0.73
	DOD-H	25	0.91	0.60	0.87	0.79	0.94	0.82
	DOD-O	55	0.90	0.52	0.86	0.74	0.92	0.79
	Mean (weighted)		0.91	0.53	0.86	0.77	0.90	0.79
	STD (weighted)		0.03	0.07	0.03	0.08	0.02	0.03

Table 9.3: Consensus score results on datasets (a) DOD-H and (b) DOD-O. Highest scores from human experts and the U-Sleep are highlighted in bold. Scores where one of the trained ML models (last 3 rows) performed as well or superior to U-Sleep are indicated by underlined numbers. However, these models were fit to the particular datasets, while U-Sleep has not seen any data from DOD-H and DOD-O during model building and training, indicated by checkmarks or crosses in the Fit column. Numbers shown are mean ± 1 standard deviation per-subject F1 scores computed between the output of a single model or human expert and the consensus scores generated from the 4 ($N-1$) remaining (when comparing human to consensus) or best human annotators (when comparing model to consensus).

(a) DOD-H: Healthy controls, $N = 25$

Scorer	Fit	Wake	N1	N2	N3	REM	Mean
Expert 1	–	0.83 ± 0.11	0.49 ± 0.15	0.86 ± 0.12	$\mathbf{0.78 \pm 0.24}$	0.84 ± 0.16	0.76 ± 0.11
Expert 2	–	0.83 ± 0.14	0.52 ± 0.11	0.88 ± 0.05	$\mathbf{0.78 \pm 0.23}$	0.89 ± 0.06	0.78 ± 0.07
Expert 3	–	0.84 ± 0.12	0.54 ± 0.13	0.88 ± 0.05	0.74 ± 0.25	$\mathbf{0.93 \pm 0.05}$	$\mathbf{0.79 \pm 0.07}$
Expert 4	–	0.73 ± 0.18	0.40 ± 0.15	0.83 ± 0.07	0.75 ± 0.22	0.90 ± 0.09	0.72 ± 0.11
Expert 5	–	0.83 ± 0.14	0.53 ± 0.12	$\mathbf{0.89 \pm 0.04}$	0.76 ± 0.24	0.90 ± 0.09	0.78 ± 0.08
U-Sleep	✗	$\mathbf{0.88 \pm 0.10}$	$\mathbf{0.56 \pm 0.14}$	0.86 ± 0.05	0.73 ± 0.23	$\mathbf{0.93 \pm 0.05}$	$\mathbf{0.79 \pm 0.06}$
SimpleNet	✓	0.83 ± 0.13	$\underline{0.57 \pm 0.14}$	$\underline{0.90 \pm 0.04}$	$\underline{0.80 \pm 0.23}$	0.90 ± 0.09	$\underline{0.80 \pm 0.07}$
DeepSleepNet	✓	0.84 ± 0.10	0.56 ± 0.13	$\underline{0.90 \pm 0.05}$	0.79 ± 0.24	0.88 ± 0.10	0.79 ± 0.07
SeqSleepNet	✓	0.81 ± 0.18	0.54 ± 0.14	0.87 ± 0.08	0.73 ± 0.25	0.86 ± 0.12	0.76 ± 0.11

(b) DOD-O: OSA patients, $N = 55$

Scorer	Fit	Wake	N1	N2	N3	REM	Mean
Expert 1	–	0.87 ± 0.11	0.38 ± 0.15	0.82 ± 0.13	0.59 ± 0.31	0.81 ± 0.25	0.69 ± 0.12
Expert 2	–	0.87 ± 0.09	0.46 ± 0.17	0.82 ± 0.11	0.61 ± 0.29	0.86 ± 0.22	0.72 ± 0.12
Expert 3	–	0.88 ± 0.09	0.42 ± 0.16	0.83 ± 0.13	0.46 ± 0.33	0.85 ± 0.22	0.69 ± 0.11
Expert 4	–	0.89 ± 0.09	0.46 ± 0.15	0.84 ± 0.07	0.52 ± 0.33	0.83 ± 0.24	0.71 ± 0.12
Expert 5	–	$\mathbf{0.90 \pm 0.08}$	0.48 ± 0.15	$\mathbf{0.86 \pm 0.08}$	0.62 ± 0.33	0.85 ± 0.22	0.74 ± 0.11
U-Sleep	✗	0.89 ± 0.09	$\mathbf{0.53 \pm 0.14}$	0.85 ± 0.08	$\mathbf{0.66 \pm 0.30}$	$\mathbf{0.88 \pm 0.20}$	$\mathbf{0.76 \pm 0.10}$
SimpleNet	✓	0.89 ± 0.09	$\underline{0.52 \pm 0.16}$	$\underline{0.88 \pm 0.11}$	0.63 ± 0.35	0.85 ± 0.22	0.75 ± 0.11
DeepSleepNet	✓	0.86 ± 0.11	0.46 ± 0.17	$\underline{0.87 \pm 0.10}$	$\underline{0.67 \pm 0.30}$	0.84 ± 0.22	0.74 ± 0.12
SeqSleepNet	✓	0.84 ± 0.13	0.46 ± 0.20	$\underline{0.86 \pm 0.10}$	0.59 ± 0.33	0.77 ± 0.28	0.71 ± 0.14

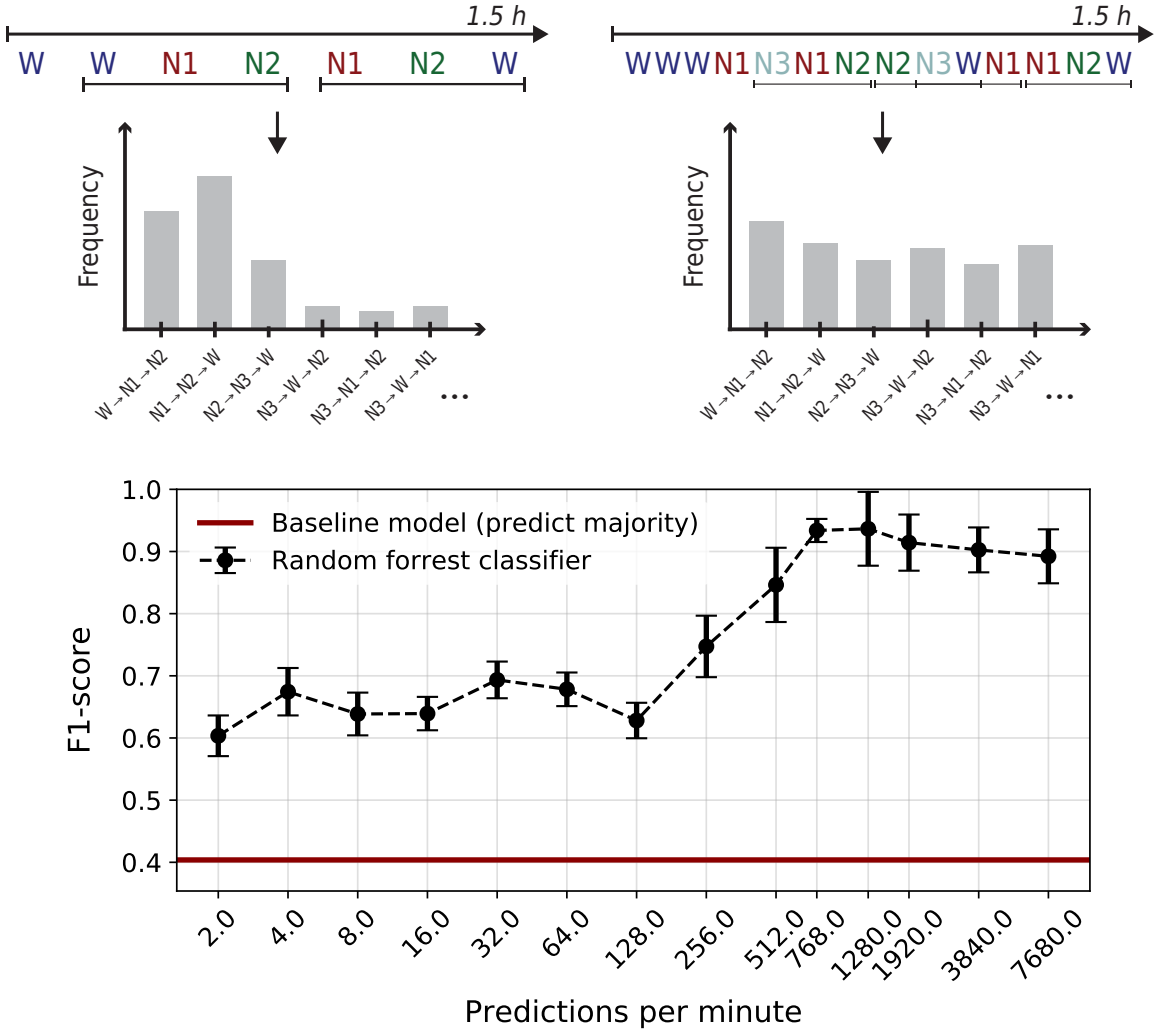


Figure 9.4: Classification performance on the task of separating healthy control subjects and OSA patients in a population of $N = 80$ (25 controls, 55 OSA patients) using a Random Forrest classifier on sleep stage transition triplet frequencies extracted using U-Sleep outputs of varying frequency. Panels (a) and (b) illustrate the process of extracting sleep stage triplet transition frequencies from low (a) and high (b) frequency outputs from U-Sleep, which are passed to the classifier. Panel (c) shows classification performance as a function of sleep staging frequency. Increasing the temporal resolution improved the predictive performance of the downstream classifier from its initial low mean F1 of 0.60 to nearly perfect classifications with mean F1 scores in range 0.89 – 0.94 at frequencies $\geq$ 768 predictions/minute. The black curve shows the mean performance with standard deviation error bars computed over 50 repetitions of the experiment using randomly configured classifiers. The solid red line is the F1 score obtained using a baseline model which predicts only the majority class (OSA patient) independent its input.

Chapter 10

Abstract E

Automatic detection of abnormal sleeping patterns in stroke patients using high-frequency sleep staging

Authors

Mathias Perslev[a], Anders Sode West[b], Sofie Amalie Simonsen[b], Laura Bødker Ponsaing[c], Helle Klingenberg Iversen[b], Christian Igel[a], and Poul Jørgen Jennum[c].

[a]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

[b]Department of Neurology, Rigshospitalet, Glostrup, Denmark

[c]Danish Center for Sleep Medicine, Rigshospitalet, Glostrup, Denmark

Published

Journal of Sleep Research 31(S1), Oral Abstract, 2022. DOI: `https://doi.org/10.1111/jsr.13739`

Introduction: Standard sleep staging methods compress polysomnographic data into 30 s sleep stages. Automatic scoring models, such as U-Sleep, make it possible to extract sleep stages at much higher frequencies. It has been hypothesized that such scores may be indicative of neurophysiological processes and may carry additional diagnostic information.

Objectives: To investigate if high-frequency outputs of the U-Sleep model contain additional diagnostic information for separating stroke patients from healthy and sleep-disordered controls compared to typical 1/30 Hz expert-derived sleep stages.

Methods: Overnight polysomnography (PSG) was performed for 20 healthy individuals, 39 patients undergoing diagnosis for sleep disorders and 233 stroke patients in the acute/sub-acute phase. Three human experts derived sleep stages at 1/30 Hz frequency, while the U-Sleep model was used to extract stages at 1/30, 1 and 12.8 Hz. Sleep stage transition-triplet frequencies were computed from expert- and automatically derived stages at each frequency. Cross-validation classification experiments using Random Forrest models were performed to separate stroke patients from healthy and sleep-disordered controls based on triplet frequency features. Classification performance was evaluated using the macro F1-score. Each experiment was repeated 50 times, and median performances obtained using automatic and expert scores were compared using two-sided Wilcoxon Signed Rank tests.

Results: U-Sleep performed sleep staging at 1/30 Hz with an F1-overlap to experts of 0.83 ± 0.18 for stage wake, 0.86 ± 0.15 for stage non-REM, and 0.64 ± 0.35 for stage REM (mean $\pm$ 1 STD, $N = 233$). Using expert derived stages, the classification experiments separated stroke patients from controls with macro F1-scores of 0.74 ± 0.01 (median $\pm$ 1 MAD, $N = 50$). In comparison, using U-Sleep scores resulted in lower classification performance at 1/30 Hz frequency (0.68 ± 0.02, $P < 0.001$), better performance at 1 Hz (0.76 ± 0.01, $P < 0.001$) and even higher performance at 12.8 Hz (0.80 ± 0.01, $P < 0.001$).

Conclusions: High-frequency sleep stage representations as output by the U-Sleep model are informative for separating stroke patients from healthy and sleep-disordered controls. Higher staging frequencies allowed for a better classification, ultimately exceeding that of human experts scores. Further work is needed to address if high-frequency U-Sleep scores reflect underlying neurophysiological processes or, for instance, model uncertainty on difficult cases.

Chapter 11

Manuscript F

U-Sleep v2: Single-channel, high-frequency, and spatial sleep staging for complex EEG

Authors

Mathias Perslev[a], Shaun Purcell[b], Poul Jørgen Jennum[c], and Christian Igel[a].

[a]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

[b]Department of Neurology, Rigshospitalet, Glostrup, Denmark

[c]Danish Center for Sleep Medicine, Rigshospitalet, Glostrup, Denmark

Published

Not yet published (under preparation).

11.1 Abstract

The U-Sleep model has been shown to perform automatic sleep staging at expert-level performance across data collected from healthy and sleep-disordered individuals. The model is robust to near-arbitrary input EEG and EOG channel combinations and has been evaluated on a large set of patient cohorts of variable demographics. We introduce an improved U-Sleep model and simplified variants requiring only single-channel input. The new models were trained on an expanded dataset of 25,805 sleep studies and were evaluated on heterogeneous testing data, including complex electroencephalography (EEG) data from patients with narcolepsy, REM Sleep Behavior Disorder, Periodic Leg Movement and Parkinson's Disease. The use of the U-Sleep model for high-frequency sleep staging was further analysed. A qualitative investigation of high-frequency scores indicated that U-Sleep detects rapid (< 30 seconds) sleep stages. On average, this effect increased the estimated duration of stages N1 and N3 sleep with a corresponding decrease in stage N2. Estimates based on high-frequency stages displayed higher correlations across repeated sleep studies of the same subjects, indicating that they may be better suited for objective sleep analysis. Finally, a study of the spatial variation of sleeping patterns when applying U-Sleep on different EEG electrodes was conducted. Spatial variation increased with increased staging frequency, with higher degrees of stage synchronisation for nearby electrodes on the same hemisphere. This paper adds to evidence supporting U-Sleep as a clinically robust model with high performance even on complex & single-channel EEG, which may be used to study sleep in novel ways in high-throughput, high-frequency studies. All models are freely available for research at `https://sleep.ai.ku.dk`.

11.2 Introduction

Sleep staging provides the basis for many clinical decisions in sleep medicine. Manual sleep staging is time-consuming and complex as human experts must evaluate hours of polysomnographic (PSG) recordings, making it expensive and subject to high inter- and intra-rater variability (Danker-Hopfe, Anderer, et al. 2009; Rosenberg et al. 2013; Younes, Kuna, et al. 2018; Younes, Raneri, et al. 2016; X. Zhang et al. 2015). Automatic sleep staging aims to automate or assist this process. The accuracy of such systems has improved dramatically in recent years with the advance of deep learning (Faust, Hagiwara, et al. 2018; Faust, Razaghi, et al. 2019; LeCun, Bengio, et al. 2015) and large data-sharing initiatives such as the National Sleep Research Resource[1] (NSRR)(G.-Q. Zhang et al. 2018) and Physionet[2] (Goldberger et al. 2000).

The development of automatic sleep stagers has been pursued since the beginning of digital sleep recordings (Bob Kemp 1993; Penzel et al. 1991; Thomas Penzel et al. 2000). A clinically robust and high-performance sleep stager is thought to improve diagnostic throughput and potentially accuracy and consistency, which may benefit patients and facilitate presumably important large-scale longitudinal sleep studies (Boostani et al. 2017; Bragazzi et al. 2019). Automatic stagers may also facilitate a transition from the traditional visual scoring of sleep physiology in 30-second discrete, whole-brain segments, which persists mainly for historical reasons and due to the complexity of visual adaptive scoring for humans (Silber et al. 2007), to a smoother and less compressed representation of sleep.

While several accurate automatic sleep staging models based on deep learning and classical machine learning techniques have been developed (Biswal, Sun, et al. 2018; Phan, Mikkelsen, et al. 2022; Supratak et al. 2017; Vallat et al. 2021), most clinical sleep staging is still performed manually into 30-second segments. Meanwhile, the recent expansion of available wearable devices for out-of-clinic and longitudinal sleeping recordings has necessitated the development of sleep stagers that work reliably with limited input data modalities.

This paper makes significant contributions within the domains of clinical sleep staging, research in alternative and more informative representations of sleeping patterns, and scoring based on single-modality recordings for, e.g., wearable devices. It shows that the U-Sleep (Perslev, Darkner, et al. 2021) model can likely support all these functions, scoring reliably on complex EEG from patient groups not seen during training, with missing input data and based on only single electroencephalography (EEG) or electrooculography (EOG) input, while also able to score sleep at higher frequencies and in distinct spatial EEG positions. The paper is divided into three parts:

[1] `https://sleepdata.org`
[2] `https://physionet.org`

Part I: U-Sleep v2 and single-channel variants The original U-Sleep model demonstrated that automatic models could be made nearly agnostic to input channel EEG and EOG derivations and that sleep staging is possible using a single input. Fiorillo, Monachino, et al. (2023) recently discovered that the original U-Sleep model is robust to even non-standard EEG and EOG derivations and could score using double EEG or EOG inputs. They also found that U-Sleep could generalize to a large and heterogeneous new dataset regardless of channel derivations used during training.

These findings imply that models like U-Sleep can be robust to variable and atypical inputs, which would be difficult for humans to score that look for specific events in specific channels when following the American Academy of Sleep Medicine (AASM) guidelines (Iber et al. 2007). It was hypothesised that U-Sleep may be well-suited for scoring complex and noisy data and that U-Sleep may provide accurate sleep scores based on only single-channel inputs for more straightforward applications in, e.g., wearable technologies.

These discoveries inspired the development of a new set of improved U-Sleep models, collectively called U-Sleep v2. These models were trained on an expanded (as compared to Perslev, Darkner, et al. 2021) dataset of 25,805 sleep studies and include a drop-in replacement to the original model, which accepts EEG and EOG inputs, as well as simplified single-channel variants. The improved U-Sleep model performed as well as or better than the original model for all stages. The new version was also shown to more accurately predict REM latency, an essential sleep metric for diagnosing narcolepsy (Sateia 2014). Although slightly below the dual-channel model, both single-channel variants performed statistically non-inferior to even the best human expert of a group of five on new data not seen during training.

Part II: Performance on complex EEG Clinically validating and implementing an automatic sleep staging system requires reliable scoring across diverse patient groups. Most research models, however, are trained on relatively homogenous patient cohorts from one or a few distinct clinical sites and evaluated on similar data, which limits the understanding of clinical robustness towards new and unexpected data which may arise in clinical practice.

U-Sleep demonstrated promising clinical stability by matching human expert performance on healthy controls and patients with obstructive sleep apnea (OSA) on data from a new sleep clinic and geographic region not seen during training. However, while U-Sleep's training and evaluation included data from over 15,000 subjects from 16 clinical studies (including both healthy individuals, sleep apnea patients and patients with various non-sleep/non-brain diseases, e.g., cardiovascular diseases), it has yet to be proven effective for several patient cohorts and complex EEG from patients with more severe brain disorders. Such studies are necessary before a complete clinical adaptation of U-Sleep can be made, as it would strongly support the general applicability of the U-Sleep model

if it were found to perform reasonably even in the most complex of cases.

The second part investigated the performance of U-Sleep v2 on several new cohorts, including people with periodic leg movements (PLM), REM sleep behaviour disorder (RBD), narcolepsy, and Parkinson's disease (PD). These experiments aimed to validate the model's clinical potential, where data may appear more complex than the model observed during training. The experiments showed that U-Sleep accurately scores sleep in narcolepsy, RBD and PLM patients and captures relevant sleep macrostructure even in severely fragmented sleep of patients with PD and RBD, several of which human annotators found difficult to score.

To further test the clinical applicability and stability of the U-Sleep model, its ability to infer sleep stages for epochs with masked data based solely on contextual information was examined. These tests simulated the impact on model stability with periods of missing input data, e.g., due to malfunctioning or paused recordings. They showed that U-Sleep often correctly scored epochs without available data using information from preceding or subsequent epochs. It was further tested how the accuracy of U-Sleep develops as input data is subjected to different band-pass filtering pre-processing and found that U-Sleep likely performs optimally with unfiltered input data. Finally, it was shown that the soft (probability-like) confidence score outputs of U-Sleep correlate with the uncertainty of a group of human experts and that these scores thus may be useful to inspect in clinical practice.

Part III: High-frequency and spatial sleep stages U-Sleep's default behaviour scores sleep stages at 1/30 Hz using all available EEG electrodes, but can also output stages at higher frequencies and score individually for each EEG electrode position, both decisive factors compared to existing validated scoring systems. This could enable the study of sleep as a near-continuous and potentially spatial physiological process rather than in whole-brain 30-second blocks.

In the original U-Sleep paper, high-frequency scores allowed for easier separation of healthy individuals from sleep-disordered patients in a statistical classification experiment, showing that the scores contained relevant clinical information. A similar result was later observed for separating acute stroke patients from the control of healthy and sleep-disordered individuals (Perslev, West, et al. 2022). Yet, whether these scores are linked to actual sleep physiology or result from a model effect on different patient groups remains unconfirmed. For instance, these experiments would have shown similar results if the high-frequency outputs were systematically noisier due to model uncertainty on the more challenging data from individuals with sleep disorders or stroke. A direct link between high-frequency stages and underlying sleep physiology should be proven to use U-Sleep to study sleep as a high-frequency physiological process.

This paper did not definitively prove such a link, but several experiments were conducted to

further the understanding of the relevance and clinical usefulness of high-frequency scores. In addition, a pilot experiment was conducted on the ability of U-Sleep to score across EEG electrode positions to study spatial sleeping patterns. These experiments were designed to complement those of Perslev, Darkner, et al. (2021) and showed that increased staging frequency affects typically derived sleep metrics, such as total sleep duration in each stage while making the estimates of such metrics more consistent across repeated studies of the same subject for most stages. Visual analysis of high-frequency scores revealed that this effect arises from U-Sleep's ability to score transient sleep stages. Finally, the synchronization of spatial sleep stages decreased with staging frequency, with spatially close electrodes on the same hemisphere supporting more similar stage predictions.

Together, this paper provides evidence in favour of adopting U-Sleep as a potential tool to enhance diagnostic throughput, scoring reliability, and robustness, facilitate out-of-clinic and longitudinal sleep studies with, e.g., wearable devices, and promote novel research in sleep physiology, transitioning towards examining sleep as a smooth, near-continuous, and possibly spatial process.

11.3 Methods

The following briefly introduces the automatic sleep staging model U-Sleep described in Perslev et al., 2021 Perslev, Darkner, et al. 2021. A set of additional datasets not considered in Perslev et al., 2019, are described, as well as how those datasets were used to train and evaluate the U-Sleep v2 model and single-channel variants. Then, the statistical evaluation of sleep staging performance is described, along with the methodological background for high-frequency and spatial sleep stages was extracted and studied.

11.3.1 U-Sleep model overview

U-Sleep is an automatic sleep staging machine learning model. It is a fully convolutional neural network (Long et al. 2014) based on the popular U-Net (Falk et al. 2019; Ronneberger et al. 2015) architecture and its time-series specialization, U-Time (Perslev, Jensen, et al. 2019). U-Sleep maps several physiological signals, such as EEG and EOG, to contiguous sleep stages, each spanning fixed-length intervals of the input signal, i.e., U-Sleep segments the input signals. U-Sleep was trained on sleep stage labels derived from human experts at $1/30\,\mathrm{Hz}$ frequency but may segment the input at higher frequencies for new data. The original U-Sleep v1 model processes exactly 1 EEG and 1 EOG input channel sampled at any frequency and is usually applied to all such available combinations to produce a majority-voted hypnogram. The input EEG and EOG channels can be recorded from a wide range of positions because the model was simultaneously trained on arbitrary combinations

of all available channel derivations over a combined training cohort of 15,660 participants and 16 clinical studies. U-Sleep v1 was as accurate as human experts on healthy individuals and OSA patients compared to the expert consensus scores (Perslev, Darkner, et al. 2021). The U-Sleep model is freely available for research at `https://sleep.ai.ku.dk`. See Perslev, Darkner, et al. (2021) for further details on the model architecture, optimization and evaluation.

11.3.2 Extended training and evaluation datasets

All training and evaluation datasets considered in Perslev, Darkner, et al. (2021), which included PSG recordings from 15,660 participants of 16 clinical studies, were considered in this study. As described in Supplementary Materials F.1, the original datasets contained atypical random channel derivations (Fiorillo, Monachino, et al. 2023), which were removed for this present study to include only typical derivations. In addition, eight new datasets were considered here for training and evaluation of the U-Sleep v2 model:

- The Nationwide Children's Hospital (NCH) Sleep DataBank (`NCHSDB` (Harlin Lee et al. 2022; G.-Q. Zhang et al. 2018), `https://sleepdata.org/datasets/nchsdb`). This database contains pediatric PSGs of 3,673 unique patients conducted at NCH in Columbus, Ohio, USA, between 2017 and 2019. This study considered 3,949 pediatric PSGs of 3,651 unique patients, as 35 recordings were excluded due to missing data, inability to safely align PSG recordings and hypnogram events, or other technical problems. Of the considered recordings, 3,794 were used to train U-Sleep v2, 53 were used to assess model performance during training, and 102 were held out for testing purposes.

- The Wisconsin Sleep Cohort (`WSC` Young et al. (2009) and G.-Q. Zhang et al. (2018), `https://sleepdata.org/datasets/wsc`). This database contains overnight in-laboratory PSGs conducted at the University of Wisconsin-Madison, Wisconsin, USA, with a sample of 1,500 state employees. A total of 2,532 recordings from 1,115 unique patients were considered. The entire database consists of 2,556 recordings, but 24 were excluded due to missing data, inability to safely align PSG recordings and hypnogram events, or other technical problems. Of the considered recordings, 2,205 were used to train U-Sleep v2, 109 were used to assess model performance during training, and 218 were held out for testing purposes.

- The Stanford Technology Analytics and Genomics in Sleep database (`STAGES` G.-Q. Zhang et al. (2018), `https://sleepdata.org/datasets/stages`). This database contains overnight in-laboratory PSGs conducted on 1,500 adult and adolescent patients at 20 data collection sites from six centres (Stanford University, Bogan Sleep Consulting, Geisinger Health, Mayo

Clinic, MedSleep, and St. Luke's Hospital). A total of 1,790 recordings from 1,559 unique patients were considered. The database consists of 2,033 recordings, but 243 were excluded due to missing data, inability to safely align PSG recordings and hypnogram events, or other technical problems. Of the considered recordings, 1,643 were used to train U-Sleep v2, 58 were used to assess model performance during training, and 89 were held out for testing purposes.

Repeated studies conducted on the same patient were treated as a single instance when assigned to either training, validation, or test-set splits not to give the system an unfair advantage. The size of the validation and testing sets was determined by selecting all recordings for up to 50 unique validation patients and 100 unique testing patients in each dataset. However, because some PSGs were excluded (e.g., due to discovered data issues) after their initial assignment to a given split, the split sizes mentioned above for each dataset may be lower.

To evaluate the sleep staging ability of U-Sleep on recordings from new patient cohorts and complex EEG patterns, an additional five datasets were considered for evaluation-only purposes:

- Danish Center for Sleep Medicine (DCSM) Narcolepsy cohort (`DCSM-N`). This dataset consists of overnight PSGs conducted on 82 unique patients with narcolepsy type I and II at the DCSM, Rigshospitalet, Denmark.

- Danish Center for Sleep Medicine (DCSM) periodic leg movement (PLM) cohort (`DCSM-PLM`). This dataset consists of overnight PSGs conducted on 41 unique patients with RBD at the DCSM, Rigshospitalet, Denmark.

- Danish Center for Sleep Medicine (DCSM) REM Sleep Behavior Disorder (RBD) cohort (`DCSM-RBD`). This dataset consists of overnight PSGs conducted on 33 unique patients with RBD at the DCSM, Rigshospitalet, Denmark. This study investigated sleep in RBD patients as well as Parkinson's Disease (PD) patients and patients with both RBD and PD diagnosed (see cohorts below).

- Danish Center for Sleep Medicine (DCSM) PD cohort (`DCSM-PD`). This dataset consists of overnight PSGs conducted on 24 unique patients with PD at DCSM, Rigshospitalet, Denmark.

- Danish Center for Sleep Medicine (DCSM) RBD & PD cohort (`DCSM-PD`). This dataset consists of overnight PSGs conducted on 31 unique patients diagnosed with RBD and PD at DCSM, Rigshospitalet, Denmark.

All PSGs of the `DCSM-N`, `DCSM-RBD`, `DCSM-PLM`, `DCSM-PD` and `DCSM-RBD-PD` datasets were used only for evaluation. None of the data was used to train any models.

11.3.3 Models

A total of 3 new versions of the U-Sleep model were developed and studied:

1. **U-Sleep v2:** A model identical to the original U-Sleep v1 model was retrained on the extended dataset described above, i.e., the same clinical cohorts that were used to train U-Sleep v1 and the training set extension cohorts of `WSC`, `STAGES` and `NCHSDB`.

2. **U-Sleep Single-Channel:** Two alternative formulations of the U-Sleep model, which process only a single EEG or single EOG input channel instead of the combined EEG and EOG input, were trained and evaluated in a similar process to that described for the U-Sleep v2 model above.

All three models were initialized with random weights with model optimization and selection conducted as described for the original model in Perslev, Darkner, et al. (2021). Most importantly, all models have trained on randomly varying input channels matching their input specifications.

11.3.4 Sleep staging evaluation

To obtain sleep stage scores for a given sleep study, the considered U-Sleep model was first applied once to all combinations of available channels that match the model's input requirements. E.g. for a study with EEG channels `{ C3-M2, C4-M1 }` and EOG channels `{ E1-M2, E2-M1 }` available, a model which requires a single EEG and a single EOG input would predict four times on the combinations `{ C3-M2 + E1-M2, C3-M2 + E2-M1, C4-M1 + E1-M2, C4-M1 + E2-M1 }`, whereas a model which requires only a single EEG input would predict on two times on `{ C3-M2, C4-M1 }`. The model confidence scores for each prediction were then summed together, and the single sleep stage that received the maximum total confidence was chosen for each epoch separately. These generated hypnograms are referred to as majority-voted model predictions. All experiments and evaluations use the majority-voted hypnograms unless explicitly stated otherwise. Only for evaluating the performance of the single-channel U-Sleep v2 variants in cases where only one channel derivation is available was the raw output from a single prediction of the U-Sleep v2 models considered.

The accuracy of a particular U-Sleep model was evaluated by comparing the macro F1/Dice score (Dice 1945; Sørensen 1948) between the majority-voted model predictions and the human expert scores. The macro F1/Dice score is the unweighted mean of stage-wise scores assigning equal importance to all sleep stages independent of the number of stage instances. Both global F1 scores computed from the summed confusion matrices across all subjects of a cohort and per-subject F1 scores (i.e., computed from the confusion matrix of a single subject) were used depending on the experiment. The former was used to simplify the analysis when large numbers of comparisons were

made as the global F1 score summarizes the performance into a single number, while the latter was used for more detailed analyses and reported with one or more summary statistics such as mean, standard deviation, median, min, max and inter-quartile range F1 scores over subjects.

Significant differences in median F1 scores between two sets of predictions on similar datasets were statistically evaluated using the paired non-parametric two-sided Wilcoxon signed rank test (`scipy` implementation, Virtanen et al. 2020) at significance level $\alpha = 0.05$. The Wilcoxon statistic (the sum of positive or negative difference ranks, whichever is smaller, denoted W) was reported along with P-values computed using the exact method, discarding pair differences that equal zero.

11.3.5 REM latency

A common failure mode for automatic sleep stagers is the wrongful prediction of so-called sleep-onset REM (SO-REM), where the atypical direct transition from Wake into REM is predicted at sleep onset. The prevalence of SO-REM predictions was quantified by calculating the REM latency calculations based on the majority-voted hypnogram outputs of the original U-Sleep v1 and new U-Sleep v2 models for all $N = 1,871$ PSGs across all validation- and test-set splits excluding datasets `WSC`, `STAGES` and `NCHSDB` to compare fairly to U-Sleep v1, which was trained without those datasets. The REM latency is the time from the first scored non-Wake (N1-N3 or REM) stage to the first scored REM stage. Note that the REM latency is 0 when a SO-REM is predicted, although here, for comparing different models, possibly wrongful predictions of (near) SO-REMs were defined as any prediction where the observed REM latency was at least 60 minutes while the predicted latency was at most 10 minutes. The prevalence of possibly wrongly predicted SO-REMs was also visually inspected in pairwise REM latency plots between the automatic and human expert-derived REM latency estimations. The REM latencies computed based on the U-Sleep v1 or v2 hypnograms were then correlated using the Pearson correlation coefficient (r). Better estimates of REM latency result in higher r values.

11.3.6 Entropy correlation to a group of human experts

U-Sleep predicts for each epoch a probability-like value for each sleep stage. When applying U-Sleep in practice, the single stage with the highest confidence is usually considered for further analysis. However, the soft probability-like values may be useful, for example, to investigate the uncertainty of the model between two or more stages in difficult or ambiguous epochs. The probability-like scores are interesting mainly if they reflect the uncertainty of a group of human expert annotators.

Whether the uncertainty of U-Sleep was correlated with the uncertainty of a group of human experts was tested using $N = 80$ PSGs from the two multi-scored evaluation-only datasets of `DOD-H`

and DOD-O. For each 30-second epoch, the stage scores of the five human expert raters were one-hot encoded and averaged to form a vector representing the overall scoring confidence of the group. The entropy of this vector was computed. Similarly, the entropy of the probability-like output vector of U-Sleep was computed for each epoch at typical staging frequency 1/30 Hz. The human and U-Sleep entropy scores were then correlated using Spearman's rank correlation coefficient (denoted ρ), a non-parametric measure of rank correlation that does not assume a linear relationship. Note that due to the distribution of human expert predictions being discrete, only seven distinct entropy values are possible given the set of 5 sleep stages, whereas the U-Sleep outputs are (up to numerical precision) continuous.

11.3.7 Robustness experiments

Two additional experiments were conducted to investigate the performance stability of the U-Sleep v2 models on input data that has been masked (to simulate missing data) or pre-process band-pass filtered.

Context awareness: Robustness to masked input data To investigate the ability of U-Sleep to sleep stage based on imperfect and noisy input data, all versions of the U-Sleep model were further evaluated on a series of $N = 39$ modified sleep studies of the test set from the dataset DCSM. As a baseline, every (overlapping) segment of signals spanning 35 epochs (each of 30 seconds duration) from all studies was scored. The centre-most prediction in each window was then extracted from each prediction and compared to the corresponding score from human experts, i.e., for each input of 35 epochs, only the prediction of the centermost epoch was considered. Predictions were made on only one channel input, i.e., majority voting was not used. The channels used for the U-Sleep v2, U-Sleep v2 (EEG) and U-Sleep v2 (EOG) models were { C4-M1, E1-M2 }, { C4-M1 } and { E1-M2 }, respectively.

Similar evaluations were made a total of 11 times but on different, noisy versions of the input data: In each repetition, one or more of the central most epoch(s) in each input segment was replaced by $N(\mu = 0.0, \sigma^2 = 0.01)$ noise in both the EEG and EOG input channels (as applicable per the model input specification), or the entire EEG or EOG channel was replaced by similar noise (only for the dual-channel model).

In one experiment, for instance, the centermost epoch of 30 seconds in each input segment of 35 epochs was replaced with noise in all input channels. All such inputs were then scored using U-Sleep, and the centralmost score was extracted from each predicted segment. This evaluation tests the ability of U-Sleep to score an epoch that contains no information in either of the input channels and must be scored solely based on its placement within signals recorded over the neighbouring epochs.

An example segment of such input data with the centermost epoch replaced by noise is visualized in Figure 11.5b. Note that only 11 out of 35 are epochs, shown in the figure for visual clarity.

A total of nine such experiments were conducted with noise replacing recorded signals in different epochs relative to the centre. These experiments included, among others, replacing the three centermost epochs with noise and replacing the epoch immediately before or after the centremost. In the most extreme case, U-Sleep had no signal information to rely on within 4 minutes and 30 seconds centred on the epoch to predict. In addition, two similar experiments were conducted, but the entire EEG and EOG channels were replaced by such noise, while the other channel contained unaltered signal data. These two experiments were conducted to test the ability of the dual-channel U-Sleep model to score only based on one of its two input channels compared to the performance of the corresponding single-channel models.

Filtering pre-processing To investigate the effect on overall U-Sleep performance with and without pre-filtering of input data, and to study the effect on stage-wise performance when filtering specific frequency bands, all $N = 36$ test-set studies of the `HPAP` dataset were scored by the U-Sleep at 1/30 Hz frequency with 16 different setups of pre-filtering applied. The `HPAP` dataset was chosen, because it consists of non-pre-filtered data. Specifically, the EEG and EOG input channels were band-pass FIR filtered using the `mne.filter.filter_data` (Gramfort et al. 2013) function with all pairs-wise combinations of parameters `l_freq` (lower pass-band edge) in [None, 0.3, 1.0, 3.0] Hz and `h_freq` (upper pass-band edge) in [None, 70.0, 35.0, 17.5] Hz, where None indicates that data is only low- or high-passed, respectively. For each of the 16 set of predictions, stage-wise and mean F1 scores were computed against the human scorer's labels to quantify the effect of the specific filtering range on both stage-wise and overall performance.

11.3.8 High-frequency sleep stage analyses

To analyse the relevancy of high-frequency U-Sleep sleep stage scores, the U-Sleep v2 model was applied to the total set of $N = 2499$ PSGs across all validation- and test-set splits with scoring frequencies 1/30 Hz (default), 1/6 Hz, 1 Hz and 5 Hz. The following experiments were conducted:

Qualitative visualizations of high-frequency scores High-frequency sleep stages were qualitatively evaluated by scoring all subjects in datasets `DOD-H` and `DOD-O` at each frequency and over one or more epochs of 30 seconds in figures exemplified by Figure 11.6a. From top to bottom, these plots visualize manual sleep scores of 5 human experts for each epoch, the consensus or majority score of the five human experts, U-Sleep's prediction at 1/30 Hz frequency, an example EEG and EOG signal and finally U-Sleep confidence scores in each sleep stage at each instance of time indicated by

stacked, coloured bar plots. These plots allow a visual analysis of how the confidence of U-Sleep in each sleep stage changes with staging frequency and how these scores align with the uncertainty of the group of human expert annotators.

Sleep stage durations To compare the total time spent in each sleep stage as estimated by U-Sleep and human experts, the durations were computed at frequencies of 1/30 Hz, 1/6 Hz, 1 Hz, and 5 Hz for U-Sleep and at typical 1/30 Hz for human experts across all $N = 2499$ from the validation- and test-set splits. Hypnograms were trimmed to include only stages between the first and last non-Wake scoring according to the scorings of the human expert. The measured durations were then compared between all sets of annotations using linear regression and the Pearson correlation coefficient (r). Significant differences in mean stage durations were statistically evaluated using the paired two-sided t-test at significance level $\alpha = 0.05$. The t statistic was reported along with P-values and Cohen's d effect size measures (Cohen 2013).

The total duration of each sleep stage was also computed within all 30-second windows based on U-Sleep's outputs at various frequencies (1/30 Hz, 1/6 Hz, 1 Hz and 5 Hz) and plotted as box plots in a confusion matrix where the conditional 30-second stage scored by human experts is represented in rows, and each stage as scored by U-Sleep in columns. These plots allow investigating if certain transient sleep stages were scored more often by U-Sleep within specific 30-second sleep stage blocks scored by humans.

Test/re-test correlations To test the hypothesis that sleep metrics derived from high-frequency sleep stage scores more accurately reflect sleep physiology, a set of test/re-test correlation experiments were conducted. Sleep stage durations were measured using outputs from U-Sleep at frequencies 1/30 Hz, 1/6 Hz, 1 Hz and 5 Hz and human experts at 1/30 Hz for all multi-visit subjects in the validation and testing splits of datasets `ABC` (11 subjects), `CHAT` (43 subjects), `MROS` (50 subjects), `NCHSDB` (5 subjects), `ISRUC-SG2` (8 subjects), `SHHS` (67 subjects), `WSC` (98 subjects) for a total of $N = 282$ unique subjects. The Pearson correlation coefficient (r) was computed between sleep stage durations measured at baseline and follow-up for each set of sleep stage annotations. Where subjects had more than two visits, all combinations of baseline and follow-up visits were considered for a total of $N = 473$ test/re-test comparisons for each sleep stage and annotator. Correlation coefficients were plotted as a function of U-Sleep staging frequency and compared to the baseline of human 1/30 Hz annotations.

Higher test/re-test correlation coefficients indicate a more robust estimate because subjects usually follow relatively similar sleeping patterns over time (at least more similar to themselves than to random other subjects in a cohort).

11.3.9 Spatial high-frequency sleep stage correlations

All U-Sleep models were trained on randomly varying input EEG/EOG (as applicable) channel combinations. Therefore, as shown in Perslev, Darkner, et al. (2021), U-Sleep is mostly invariant to channel derivations and can sleep stage given highly variable inputs. Consequently, it is possible to use U-Sleep to score sleep stages in individual channels at various physical positions. While the learned invariance properties make scorings in different input channels relatively similar at typical 1/30 Hz frequency (as shown in the Supplementary Material of Perslev, Darkner, et al. 2021), it has not been studied if spatial scoring variations appear at higher frequencies. This raises the question if it is possible to capture spatial sleep using U-Sleep by predicting individual EEG electrode positions. To study if the agreement between stages scored using different EEG input channels changes with EEG electrode position and staging frequency, the following experiment was conducted:

All PSGs from the validation- and test-set splits with all EEG and EOG channel derivations in `{ C3-M2, C4-M1, F3-M2, F4-M1, O1-M2, O2-M1, E1-M2 }` available were considered for a total of $N = 842$ PSGs. These were used to quantify spatial sleep patterns. U-Sleep v2 was then used to score each study using the six input EEG channels (each with E1-M2 as a common EOG input) separately – i.e., not using the typical majority voting scheme – at frequencies 1/30 Hz, 0.2 Hz, 1 Hz, and 5 Hz.

For each PSG, each pair of input channels and a given staging frequency, sleep staging divergence measures were computed between the probability-like confidence score outputs of U-Sleep at each epoch and averaged. Specifically, the average Jensen-Shannon divergence measure $\mathrm{JSD}(p_1 \parallel p_2) = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2}D_{\mathrm{KL}}(p_1^t \parallel m^t) + \frac{1}{2}D_{\mathrm{KL}}(p_2^t \parallel m^t)\right)$ was computed, where $m^t = \frac{1}{2}\left(p_1^t + p_2^t\right)$, D_{KL} the Kullback–Leibler divergence (Kullback et al. 1951), T the number of epochs and $p_1^t,\ p_2^t \in [0,1]^5$ the predicted softmax confidence vector at epoch t using two separate EEG channel inputs. The average divergence measure computed across all PSGs gives one similarity measure for each pair of EEG electrode positions for each staging frequency. Lower numbers indicate less divergence, i.e., more similar predictions made with each EEG input, with $\mathrm{JSD}(p_1 \parallel p_2) = 0$ only if predictions p_1 and p_2 are identical in all epochs. Higher divergence scores oppositely indicate less similar predictions with an upper bound $\mathrm{JSD}(p_1 \parallel p_2) = \log 2 \approx 0.693$ reached only if orthogonal predictions are made with each EEG input in all epochs.

The experiment was also repeated using the single-channel U-Sleep v2 (EEG) model on $N = 853$ PSGs that had all the EEG channel derivations in `{ C3-M2, C4-M1, F3-M2, F4-M1, O1-M2, O2-M1 }` available to assess the robustness of the observed patterns across two models and to investigate any possible effects on spatial patterns of using a single, shared EOG input to the dual-channel U-Sleep v2 model.

Finally, to ensure that any potential differences in scoring divergence observed at higher frequencies did not result from a fixed lag-time between pairs of signals, the JSD was also computed for all shifts of one signal relative to the other of ±60 seconds for all pair-wise EEG channels predicted at 5 Hz and plotted in a cross-correlation-like experiment. If the set of predictions in two channels were similar but offset by a fixed lag, the JSD would take its lowest value at a shift value not equal to 0.

11.4 Results

The results section contains three parts: The first describes the results of the improved U-Sleep v2 model and single-channel variants, the second their performance evaluations on complex EEG, and the third findings on high-frequency and spatial sleeping patterns.

11.4.1 Part I: U-Sleep v2 and single-channel variants

An improved U-Sleep v2 model was trained on an extended dataset (compared to Perslev, Darkner, et al. 2021) consisting of 25,805 PSGs from 16 clinical cohorts and evaluated on a total of 1,932 PSGs not used for training from 29 cohorts, 13 of which were entirely for evaluation to simulate clinical application. All models were trained using the optimization strategy described in Perslev, Darkner, et al. (2021), summarized in the Methods section. Three models were trained using different input data. The first model, U-Sleep v2, mimics the original U-Sleep v1 model, performing sleep staging using any available EEG and EOG channels. The two other models, U-Sleep v2 (EEG) and U-Sleep v2 (EOG) are simplified variants trained using only EEG or EOG input data.

U-Sleep v2 Being identical to the U-Sleep v1 model in both model definition and optimization strategy, this model serves as a drop-in replacement of the original U-Sleep model trained on an expanded dataset containing three new cohorts from the NSRR database (G.-Q. Zhang et al. 2018): The Wisconsin Sleep Cohort (`WSC`) dataset containing $N = 2532$ studies (Young et al. 2009), the Stanford Technology Analytics and Genomics in Sleep (`STAGES`) dataset containing $N = 1790$ studies, and the Nationwide Children's Hospital (NCH) and Carnegie Mellon University (`NCHSDB`) dataset containing $N = 3949$ studies (Harlin Lee et al. 2022). The cohorts are further described in the Methods section.

The best-performing model on the validation split data was found after approximately 5 million optimization steps (about 10 % longer training than U-Sleep v1) and then applied to all combinations of EEG and EOG inputs from all PSGs in the test split data. The majority vote scores across all channel combinations were compared to human expert sleep stage annotations using global (i.e., computed across all cohort subjects) and per-subject F1/Dice (Dice 1945; Sørensen 1948) score met-

rics. Table 11.1 lists these metrics for each sleep stage separately compared to similar metrics for the original U-Sleep v1 model. Performance compared to the consensus scores of 5 human experts on datasets `DOD-O` and `DOD-H` are shown in Table 11.2 and Figure 11.1. All training and evaluations were performed on a dataset with the atypical channel derivations (Fiorillo, Monachino, et al. 2023) included in the original dataset of Perslev, Darkner, et al. (2021) removed (see Supplementary Material F.1 for details).

The U-Sleep v2 model performed slightly better than U-Sleep v1 on all stages N1, N2, N3, REM and macro F1 (the unweighted mean over stage-wise scores) as measured by the weighted mean over global F1 scores, which all increased by 0.01 – 0.02 points for all. Considering the median over per-subject F1 scores, U-Sleep v2 scored significantly higher, although by similarly small absolute increases of < 0.01 – 0.02, on stages Wake, N1 N2, REM and macro F1 and indifferent on stage N3. For instance, the median ± 1 median absolute deviation (MAD) per-subject F1 REM stage score increased from 0.91 ± 0.04 to 0.92 ± 0.04 ($N = 1444$, Wilcoxon test statistic $W = 346207$, $P < 0.001$). U-Sleep v2 scored with global macro F1 scores of 0.77, 0.70 and 0.77 on the evaluation sets of the new datasets of `WSC`, `STAGES`, and `NCHSDB`, respectively. See Table 11.1 for all other statistics and comparisons to U-Sleep v1.

The U-Sleep v2 model performed similarly to the U-Sleep v1 model on the multi-scored `DOD-H` dataset ($N = 25$, mean $\pm$ 1 STD of 0.80 ± 0.08 vs 0.79 ± 0.06 and medians of 0.82 vs 0.81, $W = 93$, $P = 0.06$) and the best individual human expert (Expert 3, 0.79 ± 0.07, median 0.80, $W = 96$, $P = 0.08$). On the multi-scored `DOD-O` dataset, U-Sleep v2 scored similarly to the U-Sleep v1 model ($N = 55$, 0.76 ± 0.11 vs 0.76 ± 0.10, median 0.80 vs 0.78, $W = 741$, $P = 0.81$) and better than the best individual human expert (Expert 5, 0.74 ± 0.11, median 0.77, $W = 524$, $P = 0.04$).

The results show that the U-Sleep v2 model performed at least as well as the original U-Sleep v1 model while generalising to a larger dataset with more diverse cohorts, e.g., including more data from children (via dataset `NCDHSDB`, see Methods for details), and that U-Sleep v2 may perform better than even the best individual human expert on data from OSA patients.

Single-channel variants The single-channel EEG and EOG models were trained using the same data and optimization strategy as the U-Sleep v2 model, except only available EEG or EOG data were considered, i.e., each model was trained on random single-channel EEG or EOG inputs. The intention was to develop models better suited to settings without access to both EEG and EOG modalities (e.g., some wearable devices). The best-performing models on the validation split data were found after approximately 6.9 million and 1.2 million optimization steps (about 50 % longer and 76 % shorter training than U-Sleep v1) for the EEG and EOG models, respectively.

The performance of the U-Sleep v2 (EEG) model was compared to the dual-channel U-Sleep v2 in

Supplementary Table F.2. Channel-wise majority vote scores compared to human expert consensus scores on the datasets DOD-O and DOD-H are shown in Table 11.2 and Figure 11.1. The single-channel EEG model performed similarly to U-Sleep v2 as measured by the weighted mean over global F1 scores on stages Wake, N2, N3 and the macro F1, and lower by 0.02 points on stages N1 (0.48 vs 0.50) and REM (0.87 vs 0.89). Considering the per-subject results, U-Sleep v2 (EEG) performed significantly below, although with small absolute differences of < 0.01 – 0.02 for stages Wake, N1, N3, REM and the macro F1 ($P < 0.001$ for all) and similarly on stage N2 (0.87 ± 0.13 vs 0.78 ± 0.14, $N = 1921$, $W = 879098$, $P = 0.071$). See Supplementary Table F.2 for all other statistics and comparisons.

On the multi-scored DOD-H dataset, the single-channel model scored more accurately on average with mean 0.82 ± 0.06 vs 0.80 ± 0.08 ($N = 25$, medians 0.83 vs 0.82, $W = 89$, $P = 0.048$), while on the multi-scored DOD-O dataset the two models performed similar ($N = 55$, 0.76 ± 0.11 vs 0.76 ± 0.10, medians 0.78 vs 0.80 $P = 0.10$). U-Sleep v2 (EEG) performed better the best human rater on dataset DOD-H (Expert 3, 0.82 ± 0.06 vs 0.79 ± 0.06, medians 0.83 vs 0.80, $W = 44$, $P < 0.001$) and similar to the best human rater on dataset DOD-O (Expert 5, 0.74 ± 0.11, median 0.77, $W = 671$, $P = 0.41$).

Supplementary Table F.4 shows the performance of U-Sleep v2 (EEG) on the DCSM dataset ($N = 39$ test-split PSGs) when scoring using six individual EEG channels compared to the majority voted hypnogram generated using all channels. The table shows per-subject and per-stage statistics (mean, STD, min, max, median and IQR statistics over F1 scores). As also found in Perslev, Darkner, et al. (2021), the results show that the majority vote hypnograms are the most accurate and should be used when possible, but predictions on just a single channel may produce accurate results. For instance, U-Sleep v2 (EEG) mean ± 1 STD macro F1 scores of 0.78 ± 0.08 with median 0.80 using only the { C4-M1 } channel compared to 0.78 ± 0.09, median 0.82 for the majority voted ($N = 6$ channels) hypnograms ($W = 200$, $P = 0.01$). Refer to Supplementary Table F.4 for similar metrics on specific stages and other EEG channels.

Similar experiments and evaluations were performed for the U-Sleep v2 (EOG) model. Supplementary Table F.3 shows its majority vote performance compared to the U-Sleep v2 model. See Supplementary Material section B for detailed descriptions of these results. In summary, the EOG-only model performed slightly below the dual-channel (and the EEG-only model) model with significantly lower median per-subject scores on all stages and absolute performance differences of < 0.01 – 0.04 for all stages. However, U-Sleep v2 was still statistically non-inferior to the best of five human scorers on datasets DOD-H and DOD-O (see Table 11.2 and Figure 11.1) and scored accurately with just a single EOG channel available (see Supplementary Table F.5).

These results show that the single-channel U-Sleep v2 models are high-performance models that perform at least as well as human experts and can score accurately using a single input channel

but that the dual-channel U-Sleep v2 model should be slightly preferred when both modalities are available.

REM latency estimations REM latencies were estimated based on the majority vote scores of U-Sleep v1 and U-Sleep v2 for all 1,871 PSGs from the validation- and test-set splits – excluding the `WSC`, `NCHSDB` and `STAGES` datasets for a fair comparison to U-Sleep v1– and correlated to REM latencies observed from human scores. Figure 11.2 visualizes the experiment results. Supplementary Figure F.2 shows similar plots for the U-Sleep v2 (EEG) and U-Sleep v2 (EOG) models.

Both U-Sleep v1 and U-Sleep v2 provided accurate REM latency estimations on average with Pearson's correlation coefficients to expert observed latencies of $r = 0.82$ and $r = 0.84$, respectively. The small improvement of the U-Sleep v2 model was also reflected in fewer (likely) wrongfully predicted SO-REMs (see Methods for details) with $^{64}/_{1871}$ for U-Sleep v1 and $^{38}/_{1871}$ for U-Sleep v2. The single-channel variants similarly had slightly lower correlation coefficients compared to the U-Sleep v2 model with $r = 0.81$ and $r = 0.79$ and higher numbers of (likely) wrongfully predicted SO-REMs ($^{56}/_{1871}$ and $^{46}/_{1871}$) for the EEG and EOG model, respectively.

These results show that U-Sleep v2 has improved slightly over its predecessor to correctly separate Wake and REM near sleep onset, as also reflected in the generally slight improvement in REM stage prediction accuracy between v1 and v2 (see Table 11.1), and reiterates that while the single-channel variants are very accurate, the dual-channel model should be slightly preferred when both modalities are available.

11.4.2 Part II: Performance on complex EEG

In part I, the U-Sleep v2 and the single-channel variants were evaluated and compared on an extensive dataset of diverse cohorts of primarily healthy individuals, sleep apnea patients, and patients with various non-sleep/non-brain disorders (e.g., cardiovascular diseases). This part investigates their performance and robustness on more complex data variability to which they may be exposed in real-world clinical settings. The following experiments were conducted:

1. The scoring accuracy of the models was evaluated on five new and varied cohorts, including people with narcolepsy, PLM, RBD, PD and RBD+PD (see Methods for cohort details).

2. The correlation between the entropy (a measure of uncertainty) of the U-Sleep model and a group of human experts was measured to study if the probability-like outputs of U-Sleep may be useful in clinical practice to gauge the certainty of the model's predictions in ambiguous epochs.

3. The behaviour and accuracy of the models were studied when input was partially masked to force the models to predict based on only contextual information. This experiment simulated the response to missing data (e.g., a paused recording) or periods of insufficient data quality (to be detected by an external tool and replaced by random noise).

Additionally, the Supplementary Material F.3 contains a section on model accuracy evaluation as progressively narrower band-pass filtering is applied to input EEG and EOG data to test if U-Sleep is likely to display catastrophic performance drops if end-users pre-process their data to remove certain low- or high-frequency components. In summary, they showed that U-Sleep v2 is likely best applied to unfiltered input data, as applying even the 0.3 Hz – 35 Hz band-pass filter as recommended by the AASM guidelines (Iber et al. 2007) reduced performance, although this result may be highly dataset-specific.

Performance on complex EEG Table 11.1 shows the global F1/Dice scores on patient cohorts `DCSM-N` ($N = 82$ patients with Narcolepsy Type I and II), `DCSM-PLM` ($N = 41$ patients with PLM), `DCSM-RBD` ($N = 33$ patients with RBD, 34 PSGs), `DCSM-PD` ($N = 24$ patients with PD), and `DCSM-RBD-PD` ($N = 31$ patients with PD and RBD) all diagnosed at the Danish Center for Sleep Medicine at Rigshospitalet in Denmark.

U-Sleep v2 scored the narcolepsy cohort with a global macro F1 score similar to that of its weighted mean performance across all cohorts (0.80 vs 0.80), although with a higher than usual performance on stage Wake (0.97 vs 0.93) and lower on stages N1, N2 and REM, which may be driven by the prevalence of more Wake epochs, which are easier to score. However, the scores were generally high and well within the scores observed for other cohorts, indicating that U-Sleep can be reliably used for scoring sleep in narcolepsy patients. Similar results were observed for the PLM cohort, with average results close to the general average (0.79 vs 0.80) and slightly lower accuracies on all non-Wake stages but with high scores within the typically observed range.

The performance of U-Sleep v2 was significantly lower on the `DCSM-RBD` dataset compared to the general average (0.74 vs 0.80), and a further drop was observed on the more complex signals from patients in the `DCSM-PD` and `DCSM-RBD-PD` cohorts (macro F1 of 0.65 on `DCSM-PD` and 0.60 on `DCSM-RBD-PD`). In particular, a drop in performance was observed for stages N1 (0.29 vs a weighted average of 0.52) and REM (0.57 vs a weighted average of 0.91) for the RBD PD patients. There appeared to be an additive difficulty in scoring RBD+PD patients over patients with just RBD or PD.

Figure 11.3 and Supplementary Figures F.8–F.11 show best, median and worst case example comparisons (as measured by per-subject macro F1 scores) of hypnograms scored by U-Sleep v2 and human annotators for datasets `DCSM-PD`, `DCSM-N`, `DCSM-PLM`, `DCSM-RBD` and `DCSM-RBD-PD`, respectively.

While the scoring performance measured by F1 metrics was lower than the general average on the RBD, PD and RBD-PD patients, the scored hypnograms were visually examined and generally captured sleep macrostructure similar to that of hypnograms scored by human experts. In several cases, human annotators had difficulty confidently scoring the most complex signals from patients with severely abnormal EEG and fragmented sleep, which may affect the evaluation metrics due to ground truth uncertainty.

U-Sleep v2 performed significantly better on stage REM than U-Sleep v1 on the most difficult datasets of `DCSM-RBD`, `DCSM-PD`, and `DCSM-RBD-PD` with global F1 scores of 0.81 vs 0.74, 0.65 vs 0.57 and 0.57 vs 0.43, respectively, while the v2 model oppositely performed slightly worse on the N1, N2 and N3 stages. The lower REM performance of U-Sleep v1 on the more complex data may indicate that the original model did not fully utilize the available EOG information, which may be related to the atypical channel derivations in the original dataset (see Supplementary Materials F.1). This is supported by the single-channel variants' REM stage performances on RBD+PD, where the EOG-only model performed significantly, and atypically, better than the EEG-only model (0.53 vs 0.41), indicating that while the EEG-only model, on average, performs more accurately than the EEG-only model (see Supplementary Material F.2), the EOG channel is useful for accurate REM stage predictions when the EEG signal is complex and sleep severely fragmented. Oppositely, the EEG model performed better than the EOG model on stage N3 (0.62 vs 0.49).

The single-channel variants otherwise displayed trends similar to those of the dual-channel model, with performances on the narcolepsy and PLM cohorts similar to their typical weighted averages and gradually decreasing performance on RBD, PD and RBD+PD patients, respectively.

Entropy correlation to a group of human experts Based on visual examinations, it was hypothesised that the probability-like confidence score outputs of U-Sleep v2 correlate with the uncertainty of a group of human experts and may be useful to study in clinical practice. Figure 11.4a provides one visual example. In epochs 2 and 4, where experts disagreed on stage assignment, U-Sleep matched the experts' uncertainty assigning confidence to all stages that at least one human expert scored. The shown high-frequency stage scores (see Part III below for a detailed analysis of high-frequency stages) also reveal the potential source of uncertainty as the stage transitions from epochs 2 to 3 and 3 to 4 were more accurately detected when not restricted to 30-second epoch boundaries. Human scorers may have found assigning a single stage to the epoch difficult.

A quantification of this phenomenon is provided in Figure 11.4, which shows the median and IQR entropy values of U-Sleep v2 correlated against the entropy of 5 human expert annotators on datasets `DOD-H` and `DOD-O`. See Methods for details. The U-Sleep entropy values increased with each increase in human entropy (ignoring the likely outlier at 1.61 nats human entropy for which the sample

size was too small to make conclusions). A moderate to strong correlation (Spearman's $\rho = 0.53$, $P < 0.001$) indicates that U-Sleep and a group of experts found similar epochs easy and difficult to score unanimously. Figure 11.4 also shows that the entropy of U-Sleep v2 was, on average, higher in epochs where the group of humans scored (near) unanimously (i.e., with low entropy). However, this discrepancy may be overestimated due to the discrete set of possible entropy values that five human predictions can take. For instance, transient sleep stages in the middle of an epoch may increase the uncertainty of all human raters, yet this experiment may not capture such effects because each rater, in contrast to U-Sleep v2, which outputs soft probability-like scores, must assign a single discrete stage.

These results indicate that the uncertainty scores of U-Sleep at 1/30 Hz default scoring may be useful in clinical practice to focus the manual scorer's attention on the most difficult (as measured by disagreement between human raters) sections of the PSGs, as U-Sleep are likely to score those with high entropy values. Finally, using U-Sleep's ability to score at a higher frequency may be useful to dissect exact transition boundaries in these cases. See Part III below for further details.

Robustness to masked input data The robustness of the U-Sleep v2 models towards partially masked input data was studied by replacing sub-sections of input data with random noise and comparing their performance to when unaltered, full-information inputs were scored.

Across $N = 39$ sleep studies in the test set split of the dataset `DCSM`, 59.084 overlapping segments of 35 epochs of signals were processed by all models once for each noise experiment. Each experiment's prediction for these central epochs was compared with the human expert labels, and stage-wise F1 scores were computed for each sleep study separately. See the Methods section for setup for additional details.

Figure 11.5 visualizes F1 scores for the U-Sleep v2 model as boxplots for several experiments with different lengths of input signal replaced by noise. Similar plots for U-Sleep v1, U-Sleep v2 (EEG) and U-Sleep v2 (EOG) are shown in Supplementary Figure F.3. As expected, all models performed best with full-information inputs and saw performance drops as input signal noise increased. The re-trained U-Sleep v2 model experienced a significant decrease in mean F1 scores when replacing the entire EEG or EOG input with noise from 0.77 ± 0.08 (median 0.81) to 0.74 ± 0.10 (median 0.78, $W = 112$, $P < 0.001$) and 0.74 ± 0.09 (median 0.75, $W = 78$, $P < 0.001$), respectively. For comparison, the U-Sleep v2 (EEG) scored 0.76 ± 0.09 (median 0.78, $W = 258$, $P = 0.07$), and the U-Sleep v2 (EOG) scored 0.74 ± 0.10 (median 0.78, $W = 336$, $P = 0.46$) in their full-information setting (i.e., no input data replaced by noise), indicating that U-Sleep v2 on this dataset performed as well as the single-channel variants when exposed only to one of its two expected inputs. Removing EOG for the U-Sleep v2 model had a more significant negative impact on REM stage accuracy (from

0.89 ± 0.11 to 0.81 ± 0.18) while removing EEG affected N1 stage performance more (from 0.52 ± 0.12 to 0.48 ± 0.13).

When replacing data in both input channels of the central epoch with noise for the U-Sleep v2 model, mean performance dropped from 0.77 ± 0.08 (median 0.81) to 0.71 ± 0.09 (median 0.71, $W = 13$, $P < 0.001$). This was mainly driven by decreased stage N1 performance from 0.52 ± 0.12 (median 0.54) to 0.41 ± 0.10 (median 0.42, $W = 50$, $P < 0.001$). Other stages showed less, although statistically significant, decline, e.g., the REM stage from 0.89 ± 0.11 (median 0.92) to 0.88 ± 0.13 (median 0.91, $W = 197$, $P = 0.01$). Similar patterns occurred when removing epochs before/after the central epoch, with N1 performance dropping more significantly than the remaining stages. This is likely related to N1 being short and transitive, while other stages can more often be inferred from the surrounding context alone. One example where a transient N1 stage was correctly predicted from context is visualized in Figure 11.5b. Figure 11.5c shows an example where U-Sleep failed to infer the central epoch without data, although, in this example, it predicted a reasonable transition from Wake to N2 via N1, assigning probabilities to both bordering stages despite the lack of central epoch data.

Interestingly, N1 performance was more negatively affected by removing the central and two preceding and proceeding epochs (0.29 ± 0.11, median 0.30) than by removing the central and four preceding (0.36 ± 0.10, median 0.36, $W = 86$, $P < 0.001$) or proceeding (0.40 ± 0.11, median 0.42, $W = 30$, $P < 0.001$) epochs. This suggests U-Sleep can sometimes infer the central stage for N1 based on pre-epoch (preferred) or post-epoch information alone but requires information from the immediate context to detect the often transient stage.

Additional context visualization examples are shown in Supplementary Figures F.4-F.7, all of which show different noise settings and indications of how U-Sleep successfully and unsuccessfully relied on neighbourhood and/or sleep stage transition pattern probabilities to infer stages for epochs with missing data.

11.4.3 Part III: High-frequency and spatial sleep stages

In Parts I and II, all experiments considered sleep staging at the default scoring frequency of 1/30 Hz and most experiments considered the majority-voted hypnogram computed across all available channel inputs. To further the understanding of the (potential) physiological and clinical relevance of U-Sleep's high-frequency and spatial sleep stage scores, the following experiments were conducted:

1. The effect of scoring at higher frequencies on derived metrics such as total sleep duration in each stage was investigated, and the robustness of such metrics as a function of staging frequency was quantified with test/re-test experiments.

2. A pilot study investigated the similarity of stages predicted in different spatial EEG locations depending on staging frequency.

Figure 11.6a shows a qualitative example of why high-frequency stages may be interesting to study. The U-Sleep v2 model was applied to data from a sleep study from the `DOD-H` dataset. All five human annotators and the U-Sleep model consistently scored the third and fourth epochs as stage Wake. At 1/30 Hz frequency, the U-Sleep model indicates a near 100 % confidence in stage Wake for both epochs. However, there is a visually apparent dampening in the signal amplitude of about 15 seconds between the two epochs, which could indicate a rapid physiologically relevant change in sleep, which is not reflected by the typical 30-second annotations. At higher scoring frequencies, U-Sleep detects a rapid transition from Wake to N1 and/or N2 before regaining confidence in stage Wake. As addressed in detail in the Discussion section below, it is difficult to conclude if U-Sleep high-frequency confidence scores reflect actual sleep physiology. However, examples such as this are abundant (see Supplementary Materials) and raise whether high-frequency scores capture rapid sleep stage transitions lost at the 30-second scale. The following experiments aimed to quantify these and other observed effects of high-frequency staging.

Sleep stage durations & test/re-test correlations Table 11.3 (see also Supplementary Figure F.13 for visual representations of the table values) and Supplementary Figure F.12 quantify the effect of transient sleep stages that may appear within 30-second epochs. Table 11.3 shows the average total durations of each sleep stage as measured using human-derived annotations and U-Sleep v2 at frequencies 1/30 Hz, 0.2 Hz, 1 Hz and 5 Hz across 2499 PSGs from the validation- and test-set splits. It also shows test/re-test Pearson's correlation values (r) comparing stage duration estimates computed at different longitudinal visits of the same patient, higher values of which may indicate a more robust duration estimator (see Methods for details and argumentation). A correlation coefficient was computed for each duration estimate (i.e., for all scorings and stages) across 473 total test/re-test comparisons of 282 unique patients.

There were only minor differences between the mean durations measured using human-derived annotations and U-Sleep-derived annotations (at 1/30 Hz) for all stages W (1.42 ± 1.51 vs 1.41 ± 1.45, $N = 2499$, $t = -1.42$, $P = 0.15$, Cohen's $d = -0.01$), N2 (3.43 ± 1.06 vs 3.54 ± 0.96, $t = -9.14$, $d = -0.11$), N3 (1.08 ± 0.88 vs 1.05 ± 0.77, $t = 3.39$, $d = 0.04$) and REM (1.18 ± 0.58 vs 1.22 ± 0.57, $t = -7.54$, $d = -0.08$). While all differences were significant according to the paired t-test, this was mainly due to the large paired sample size, as the effect sizes (Cohen's d) were all small (< 0.2, Cohen 2013). Human annotators, however, scored a higher mean N1 duration than U-Sleep at 1/30 Hz (0.60 ± 0.50 hours vs 0.47 ± 0.38 hours, corresponding to an ≈ 28 % average difference, $t = 18.13$, $P < 0.001$, $d = 0.30$), but with lower test/re-test correlation as compared to U-Sleep at

1/30 Hz (0.55 vs 0.66). When the staging frequency increased, U-Sleep measured increasingly higher N1 mean durations (from 0.47 to 0.66) and with higher test/re-test correlations (from 0.66 to 0.70). These results show that U-Sleep scored total N1 sleep duration more consistently between baseline- and follow-up visits than human annotators, with further improved consistency when scoring at higher staging frequencies. Note that higher test/re-test correlations do not alone indicate that the duration estimates are more aligned with the true physiology, only that the estimates are more reliable across visits. The predicted N1 duration, for instance, starts lower (mean 0.47 hours) and ends higher (0.66 hours) than the human-scored durations (0.60), yet the correlation values were higher for U-Sleep in both cases.

Similarly, average durations measured by U-Sleep changed with staging frequency for stages Wake (from 1.42 hours to 1.56 hours, $\approx$ 10 % increase), N2 (from 3.54 hours to 3.01 hours, $\approx$ 15 % decrease), N3 (from 1.05 hours to 1.29 hours, $\approx$ 23 % increase). The test/re-test correlation increased slightly for all of the stages Wake (0.45 to 0.46), N2 (0.48 to 0.52) and N3 (0.79 to 0.82). The test/re-test correlation was higher for U-Sleep, even at 1/30 Hz frequency, compared to human annotators for all stages: Wake, N1, N2 and N3. Oppositely, neither the measured total duration (from 1.22 hours to 1.18 hours, $\approx$ 3 % decrease) nor r value (0.37 to 0.37) changed notably for stage REM with changing frequency, and the r value was consistently lower for U-Sleep as compared to human annotators.

Figure 11.6b plots durations measured using U-Sleep at 1/30 Hz against human annotations. The plot shows that while both sets of annotations provided linearly correlated duration measurements, particularly for stages Wake, N2, N3, and REM, there were several individual significant exceptions where one annotator scored significantly shorter total duration for that stage. For all stages, there are cases where U-Sleep at 1/30 Hz scores a particular stage significantly less often and significantly more frequently than the human annotator. Figure 11.6c shows a similar plot, but which compares durations using annotations from U-Sleep at 1/30 Hz with U-Sleep at 5 Hz. A near-perfect correlation was found between sleep durations measured at 1/30 Hz and 5 Hz for stages Wake and REM ($r = 0.99$ and $r = 0.99$) with no significant outliers (i.e., dots away from the identity line). In contrast, as also visible from Table 11.3, measured N2 stage duration was consistently lower at 5 Hz compared to 1/30 Hz, with oppositely higher estimated N1 and N3 durations at 5 Hz. In combination with the results of Supplementary Figure F.12, which shows the total confidence of the U-Sleep model in each possible class within windows of 30-seconds conditioned on human annotation for that interval, these results show that U-Sleep tends to, in particular, score more transient N1 and N3 stages within periods that would be scored N2 by both humans and U-Sleep at typical 1/30 Hz.

Spatial high-frequency sleep stages Because U-Sleep can score in any input channel, it was hypothesised that spatial sleep dynamics might be detected by applying U-Sleep to individual EEG

electrode positions. Figure 11.7 shows the results of a pilot study of such spatial high-frequency sleep staging patterns. The mean Jensen–Shannon divergence (JSD) measure was computed between pairs of U-Sleep v2 predicted confidence scores in different spatial EEG electrode positions and at different frequencies (1/30 Hz, 1/6 Hz, 1 Hz and 5 Hz) across $N = 842$ validation- and test-set PSGs that had had all of EEG channels `{ C3-M2, C4-M1, F3-M2, F4-M1, O1-M2, O2-M1 }` and EOG channel E1-M2 available. See Methods for further quantification details.

Figure 11.7a shows that the similarity of predictions made at typical 1/30 Hz in any pair of EEG channels was high (i.e., low mean JSD) with scores close zero (≤ 0.022 for all pairs, mean 0.015, range $[0, \log 2 \approx 0.693]$). In other words, the confidence score outputs of U-Sleep were similar in all EEG channels. This result aligns with the findings of Perslev, Darkner, et al. (2021) where the (thresholded/discrete) scoring performance was similarly high using any EEG channel. These results were expected and are a consequence of the induced channel invariance of the U-Sleep training pipeline (see Perslev, Darkner, et al. 2021 for details).

Figures 11.7b-11.7d show the spatial scoring divergences as the scoring frequency was increased. Notably, the divergence increased between all pairs of EEG channels with increased frequency with a mean JSD of 0.022 at 1/6 Hz, 0.041 at 1 Hz, and 0.077 at 5 Hz, i.e., an average JSD approximately 5 times higher at 5 Hz than at 1/30 Hz (the absolute divergences remain, however, low). Supplementary Figure F.15 shows two examples of the cross-correlation-like experiment, described further in the Methods section, which showed that the increasing divergence observed at high scoring frequency is not a simple result of a linear, fixed lag between the compared time-series, as the lowest JSD was observed in nearly all PSGs and channel combinations at offset 0 (i.e., no relative shift).

At 5 Hz, where the spatial scoring divergence (dis-similarity) was highest, the most similar predictions were made in channel pairs `{ F3, C3 }` and `{ F4, C4 }` with mean JSD scores of 0.042 and 0.039, respectively. The highest divergence was observed between pairs `{ O2, F3 }` and `{ O1, F4 }` with mean JSD scores of 0.100 in both cases.

The relatively closely positioned – but contralateral – occipital electrodes (O1 and O2) were scored with a mean JSD of 0.084, which is slightly larger than the divergences between the occipital electrodes and their same-hemisphere, but relatively far apart central counterparts (C3 and C4) with mean JSD scores of 0.070 for channel pair `{ O1, C3 }` and 0.068 for channel pair `{ O2, C4 }`. These results show that U-Sleep v2 scored more similarly at high frequency in nearby EEG electrodes on the same hemisphere. In contrast, more dis-dissimilar predictions were made in EEG electrodes placed contralaterally and far apart.

The experiment was repeated using the single-channel U-Sleep v2 (EEG) model to rule out any potential bias effects of using a common `{ E1-M2 }` EOG input to the U-Sleep v2 model above. Similar patterns were observed with slightly higher average divergence scores. See Supplementary

Figure F.14. These results suggest that the shared EOG input makes the U-Sleep v2 score more similarly across EEG inputs, but that spatial EEG variations drive the spatial scoring patterns observed at high-frequency outputs.

11.5 Discussion

This paper introduced new and improved U-Sleep v2 models and stress-tested them on several parameters to advance the understanding of their performance and behaviour in clinical practice. The models were trained on an extensive and heterogeneous multi-site dataset of 25,805 PSGs and evaluated on complex EEG from patients not seen during training to simulate a real-world application. Single-channel variants were developed for easier integration with wearable devices. The clinical applicability and stability of the models were examined through various scenarios, including masked input data, different band-pass filtering, and correlation with human expert uncertainty. The study also investigated the clinical relevance of high-frequency sleep stage outputs and explored spatial sleeping patterns using U-Sleep's special scoring capabilities in a pilot study.

Three variants of U-Sleep v2 were introduced: U-Sleep v2, which is a drop-in replacement to U-Sleep v1, accepting both EEG and EOG inputs, was found at least as accurate as the original model while generalizing to an even greater dataset, including, for instance, additional pediatric sleep studies. The U-Sleep v2 model also improved over its predecessor on a common failure mode for automatic sleep staging models, which is the wrongful prediction of SO-REMs. It predicted fewer likely wrong instances and obtained a higher SO-REM estimation correlation to human experts.

Two single-channel alternatives, U-Sleep v2 (EEG) and U-Sleep v2 (EOG), were also developed and assessed. The single-channel variants demonstrated slightly lower performance than the average of the U-Sleep v2 model, yet they still exhibited high performance, even in extreme cases where only one input channel was available. U-Sleep v2 (EEG) even outperformed the dual-channel model on the multi-scored `DOD-H` dataset, and both single-channel EEG and EOG models were statistically non-inferior to the best human annotator on both the `DOD-H` and `DOD-O` datasets.

The U-Sleep v2 models were assessed on several patient groups not well-represented in the training dataset. U-Sleep v2 accurately scored sleep for narcolepsy and PLM patients with a performance similar to that observed for healthy individuals and OSA patients. The model's performance was slightly lower for RBD patients; however, it was not conclusive whether the reduced scoring accuracy was causally linked to the RBD disorder cohort factor. The average performance (mean F1 of 0.74) remained high and fell within the range of scores observed in other cohorts, such as `STAGES` (0.70), `MASS-C1` (0.72), and `SVUH` (0.73). The reason for the lower performance in these specific cohorts is still unknown. U-Sleep v2 was further evaluated on complex EEG patterns from PD and RBD+PD

patients. Although the model's performance was significantly lower than average for these cohorts, visual inspections of the resulting hypnograms indicated a good representation of the overall macro sleep structure. Fragmented sleeping patterns made objective, quantitative evaluation challenging due to unstable evaluation metrics and uncertainty in the ground truth, as even human annotators found several PSGs difficult to score.

U-Sleep v2 (EEG) demonstrated approximately similar performance to the dual-channel model when scoring narcolepsy, PLM, and PD patients, with mean F1 scores of 0.79 vs 0.80, 0.78 vs 0.79, and 0.66 vs 0.65, respectively. It scored RBD and RBD-PD patients slightly less accurately with global macro F1 score differences of ≈ 0.02 points. Interestingly, the single-channel EEG model showed significantly lower accuracy in scoring REM sleep for the RBD (0.76 vs 0.81) and RBD+PD cohorts (0.44 vs 0.57) but performed slightly better on REM for the PD cohort (0.67 vs 0.65, respectively). This suggests that the EOG channel provides significant information to correctly score REM sleep (likely by enhancing the model's ability to distinguish REM sleep from Wake in cases of highly fragmented sleep) but is required only in certain complex cases. This is supported by the U-Sleep v2 (EOG) results, which scored REM sleep significantly better than U-Sleep v2 (EEG) in RBD+PD patients, although performing slightly worse on average on most other cohorts.

The above results indicate that the U-Sleep v2 models exhibit high average performance and can effectively generalize to even complex new patient cohorts not encountered during training. However, a thorough analysis of the generated hypnogram should be conducted in the most complex cases involving severe sleep fragmentation or REM sleep abnormalities, e.g., in patients with neurodegenerative disorders or RBD. The single-channel U-Sleep v2 variants performed as well as the best human experts on healthy and OSA patients not seen during training and could score accurately using a single input electrode. The single-channel U-Sleep models may thus be effective for use in wearable devices to support, for instance, longitudinal studies. However, when both EEG and EOG are available, the dual-channel U-Sleep v2 model should be marginally preferred for scoring complex cases with maximum accuracy and consistency.

The entropy (a measure of uncertainty) of the outputs of U-Sleep v2 at a typical $1/30\,\mathrm{Hz}$ scoring frequency was found to correlate with the uncertainty of a group of five human expert scorers. Visual studies of high-frequency scores (discussed further below) revealed that uncertainty might arise in boundary regions with miss-aligned epoch boundaries where unanimously assigning a single stage to a 30-second epoch is non-trivial. These results indicate that U-Sleep's confidence outputs may be useful in clinical practice to direct the attention of the human scorer to the most relevant, difficult parts of the recording. High-frequency scores can support this process by allowing human operators to evaluate possible sources of uncertainty (such as a transient stage).

The performance of the U-Sleep v2 model was further investigated under the influence of masked input data by replacing parts of input data with random values to simulate malfunctioning or paused recordings or otherwise missing data. As anticipated, scoring accuracy decreased when input data was missing. However, using contextual information, U-Sleep could often accurately infer sleep stages for epochs with no available data. It was also discovered that, at least on the `DCSM` dataset, the U-Sleep v2 model performed approximately on par with the single-channel variants when provided with only one input and noise as the second input.

It is, however, worth noting that U-Sleep's ability to perform with input data replaced by noise is expected, as the models were trained with random data replacements for data augmentation to force the model to learn long-range dependencies. Introducing other types of noise or artefacts may have a more negative impact on performance, and it remains uncertain whether, in the case of true noise or malfunctioning recordings, input data should be replaced by the type of noise used for augmentation during training or if the model should be exposed to the raw signal containing its natural noise. The filtering experiments indicated that U-Sleep is likely best applied to relatively raw (in this case unfiltered) input data for optimal performance, as even applying the AASM recommended 0.3 Hz – 35 Hz band-pass filtering reduced the model's performance on the `HPAP` dataset. However, this result may be dataset-specific and necessitates further experimentation. The U-Sleep v2 model was trained on a subset of data from the `HPAP` dataset, which was not pre-filtered, potentially leading to a bias towards higher performance on non-filtered data for this specific dataset.

The significance of U-Sleep's high-frequency and spatial sleep stages was examined in the third and final part of the paper. Visual inspections suggested that high-frequency outputs might enable more accurate quantification of transient sleep stages and stage transitions that do not align with the arbitrary 30-second epoch boundaries. With the finding that U-Sleep's scoring uncertainty correlates with that of human experts, discussed above, these findings imply that the high-frequency and probability-like outputs of U-Sleep offer an interesting alternative representation of sleep stages, potentially containing more information than traditional discrete stages.

It was demonstrated that the U-Sleep v2 model, even at the typical 1/30 Hz frequency, provided sleep stage duration estimates with high correlation to those of human experts while offering more consistent scoring across multiple visits of the same patients for stages Wake, N1, N2, and N3. However, human annotators scored the REM stage more consistently. The consistency of stage duration estimates generally increased as the U-Sleep scoring frequency increased (except for REM, which remained stable). This could suggest that U-Sleep's high-frequency outputs are clinically relevant, serving as a more reliable objective measure for some sleep metrics. However, this result may be metric-specific. As observed in high-frequency plots such as Figures 11.5, 11.4 and Supplementary

Figures F.4–F.7, high-frequency scores can be noisy or highly responsive to local signal patterns. Some metrics, like REM latency, may require a new definition based on robust statistics rather than the current outlier and noise-sensitive definition of time from the first non-Wake to the first REM stage. Different metrics might also be calculated based on various frequency outputs from U-Sleep as applicable.

Interestingly, the average stage durations also varied with scoring frequency, increasing by approximately 10 % for stage Wake, 28 % for stage N1, and 23 % for stage N3, while the average stage N2 duration decreased by around 15 %. A possible explanation for the non-REM variations is that U-Sleep can score transient N1 and N3 stage sleep at higher frequencies when not constrained by epoch boundaries and not required to assign a single stage to epochs with multiple sub-stages. In such cases, e.g., where a 30-second period has characteristics of multiple non-REM stages, epochs might otherwise tend to be scored as the middle-most and majority stage, N2.

In the final analysis, this study investigated if sleep scoring varies with the spatial location of input EEG electrodes. Since U-Sleep is trained to induce invariance to the exact electrode positions, the experiments predictably demonstrated high-scoring similarity across electrodes at a typical 1/30 Hz frequency. However, as the scoring frequency increased, so did the scoring variability with spatial EEG location. Predictions made using nearby electrodes were more similar than those between distant electrodes. A weak tendency for more similar scoring between electrodes placed on the same hemisphere was also observed. Cross-correlation-like experiments showed that a fixed lag-time effect did not cause the observed drop in scoring similarity at higher stage frequencies. These results suggest that U-Sleep could potentially be used to study sleep as a local phenomenon, which may vary in health and disease (although this remains to be demonstrated).

Establishing a direct causal relationship between the high-frequency or spatial U-Sleep scores and underlying sleep physiology remains challenging. However, Perslev, Darkner, et al. (2021) demonstrated that high-frequency stages enable easier separation of healthy control subjects and OSA patients, which was later reaffirmed for separating stroke patients from a control group of healthy and OSA patients. These findings reveal that some information in the high-frequency scores is compressed and lost in the typical 1/30 Hz scoring regime. Based on visual analysis and the computation of sleep stage durations at different frequencies, this paper suggests that U-Sleep may be capable of inferring transient sleep stages (i.e., those lasting less than 30 seconds) and that some sleep metrics can be more robustly estimated at higher scoring frequencies. The pilot experiment on spatial sleep staging indicated that high-frequency stages systematically vary with spatial EEG location. These results further suggest informative content and provide an early indication of a, but importantly not proven, causal relationship between high-frequency and spatial stages and sleep physiology.

In combination, this paper introduced new U-Sleep models and presented further evidence sup-

porting their suitability for clinical sleep staging. U-Sleep can accurately score healthy and sleep-disordered individuals and provide reasonable predictions even on highly complex EEG data from subjects with brain disorders like PD without being trained on such data. The models are also robust, capable of scoring using any input EEG or EOG channels and even just a single channel, making them suitable for wearable or home-testing setups. Lastly, the high-frequency outputs and ability to score spatial sleep lay the groundwork for computing new and more robust clinical sleep metrics and open up opportunities for novel research in basic sleep physiology.

Ethical Approval

The Research Ethics Committee for SCIENCE and HEALTH, University of Copenhagen, has reviewed this research project and has found it compliant with the relevant Danish and International standards and guidelines for research ethics. All `DCSM` datasets were extracted and anonymized by the Danish Center for Sleep Medicine under general approval from the Danish Data Protection Agency to analyze historical PSG data. All other datasets were acquired from third-party databases and handled according to the relevant data-sharing agreements.

Code Availability

The in-house developed codebase for training U-Sleep is publicly available on GitHub at `https://github.com/perslev/U-Time`. The codebase is supplied with the submitted manuscript. The software includes a command-line interface for initializing, training and evaluating models without altering the underlying codebase. The software is based on `TensorFlow` (Abadi et al. 2015). Please refer to the `README` file of the repository for guidance on installation and a step-by-step guide on how to train a U-Sleep model.

We make inferences using all pre-trained U-Sleep models freely available at `https://sleep.ai.ku.dk` for non-commercial usage. Users may upload (anonymised or public domain) PSG files (European Data Format, EDF, or HDF5) to the service, choose parameters such as which channels to use and the inference frequency (e.g., 1/30 Hz or higher), and receive back the automatically scored hypnogram. The service also provides a simple interface to interactively visualize the scored hypnogram and obtain key sleep statistics over selected periods. The raw sleep stages can be downloaded in several formats. An API is made available free of charge for programmatic access to the web service. See `https://sleep.ai.ku.dk/docs/api/overview`. Python bindings and a command-line API interface are available at `https://github.com/perslev/U-Sleep-API-Python-Bindings`.

Acknowledgements

We gratefully acknowledge support from the Independent Research Fund Denmark through the project „U-Sleep“ (project number 9131-00099B). We also would like to thank the BETA.HEALTH organization (`https://betahealth.dk/`) for supporting the upcoming clinical validation study and implementation of U-Sleep. Finally, we thank many great colleagues from the Danish Center for Sleep Medicine, Rigshospitalet, Denmark, for their support, data collection, and curation.

The National Sleep Research Resource was supported by the U.S. National Institutes of Health, National Heart Lung and Blood Institute (R24 HL114473, 75N92019R002).

NCH Sleep DataBank was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Number R01EB025018. The National Sleep Research Resource was supported by the U.S. National Institutes of Health, National Heart Lung and Blood Institute (R24 HL114473, 75N92019R002).

This research has been conducted using the STAGES - Stanford Technology, Analytics and Genomics in Sleep Resource funded by the Klarman Family Foundation. The investigators of the STAGES study contributed to the design and implementation of the STAGES cohort and/or provided data and/or collected biospecimens, but did not necessarily participate in the analysis or writing of this report. The full list of STAGES investigators can be found at the project website.

The Apnea, Bariatric surgery, and CPAP study (ABC Study) was supported by National Institutes of Health grants R01HL106410 and K24HL127307. Philips Respironics donated the CPAP machines and supplies used in the perioperative period for patients undergoing bariatric surgery.

The Cleveland Children's Sleep and Health Study (CCSHS) was supported by grants from the National Institutes of Health (RO1HL60957, K23 HL04426, RO1 NR02707, M01 Rrmpd0380-39).

The Cleveland Family Study (CFS) was supported by grants from the National Institutes of Health (HL46380, M01 RR00080-39, T32-HL07567, RO1-46380).

The Childhood Adenotonsillectomy Trial (CHAT) was supported by the National Institutes of Health (HL083075, HL083129, UL1-RR-024134, UL1 RR024989).

The Home Positive Airway Pressure study (HomePAP) was supported by the American Sleep Medicine Foundation 38-PM-07 Grant: Portable Monitoring for the Diagnosis and Management of OSA.

The Multi-Ethnic Study of Atherosclerosis (MESA) Sleep Ancillary study was funded by NIH-NHLBI Association of Sleep Disorders with Cardiovascular Health Across Ethnic Groups (RO1 HL098433). MESA is supported by NHLBI funded contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168 and N01-HC-95169 from the National Heart, Lung, and

Blood Institute, and by cooperative agreements UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420 funded by NCATS.

The National Heart, Lung, and Blood Institute provided funding for the ancillary MrOS Sleep Study, "Outcomes of Sleep Disorders in Older Men," under the following grant numbers: R01 HL071194, R01 HL070848, R01 HL070847, R01 HL070842, R01 HL070841, R01 HL070837, R01 HL070838, and R01 HL070839.

The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University).

The Study of Osteoporotic Fractures (SOF) was supported by National Institutes of Health grants (AG021918, AG026720, AG05394, AG05407, AG08415, AR35582, AR35583, AR35584, RO1 AG005407, R01 AG027576-22, 2 R01 AG005394-22A1, 2 RO1 AG027574-22A1, HL40489, T32 AG000212-14).

The ISRUC-SLEEP Dataset has been supported by the Portuguese Foundation for Science and Technology (FCT) under Ph.D. Grants SFRH/BD/81828/2011 and SFRH/BD/80735/2011 and by QRENfunded project SLEEPTIGHT, with FEDER reference CENTRO-01-0202-FEDER-011530. We would also like to acknowledge sleep experts from the Sleep Medicine Centre of the Hospital of Coimbra University (CHUC), for their invaluable support in data acquisition and visually scoring the PSG recordings of ISRUC-Sleep dataset.

Author Contributions

All authors contributed to the design of the study; M.P. and C.I. developed the deep learning system. M.P. implemented the system and conducted the experiments; M.P. wrote the paper with continuous feedback from C.I., P.J.J. and S.P.; All authors approved the final manuscript.

Table 11.1: Channel-wise majority vote sleep staging performance comparisons of U-Sleep v1 and U-Sleep v2. The models were compared across all test-set splits using global (i.e., computed across all subjects in a cohort) F1/Dice scores. The *macro* column is the macro F1 score (i.e., unweighted mean over the global stage-wise F1 scores). Scores of U-Sleep v1 and U-Sleep v2 are shown to the left and right of each table cell, respectively, with an arrow (→) in between unless both models archived a similar score. A bold font indicates that U-Sleep v2 scored better, and an underlined font indicates that U-Sleep v1 scored better. Below the main table are summary statistics comparing the sample weighted mean and standard deviations of both models computed over the global F1 scores as well as median, median absolute deviation (MAD), and Wilcoxon's test statistics computed over *per-subject* F1 scores (i.e., in contrast to the global F1 scores of the upper table) for comparison. U-Sleep v2 was also evaluated on the test-splits of datasets `WSC`, `STAGES` and `NCHSDB`, but scores on those datasets were not included in the summary statistics and tests to allow direct comparison to the U-Sleep v1 model.

Type	Dataset	Records	Wake	N1	N2	N3	REM	Mean
Internal - Train/Test	ABC	20	**0.89 → 0.90**	**0.42 → 0.62**	**0.83 → 0.84**	0.73	**0.91 → 0.92**	**0.76 → 0.80**
	CCSHS	78	0.97	**0.61 → 0.64**	**0.91 → 0.92**	**0.87 → 0.89**	**0.92 → 0.93**	**0.86 → 0.87**
	CFS	92	0.96	**0.52 → 0.53**	**0.88 → 0.89**	**0.81 → 0.86**	**0.90 → 0.92**	**0.81 → 0.83**
	CHAT	128	**0.96 → 0.97**	**0.60 → 0.63**	**0.85 → 0.87**	**0.88 → 0.90**	**0.89 → 0.91**	**0.84 → 0.85**
	DCSM	39	**0.98 → 0.99**	**0.47 → 0.55**	0.86	0.83	**0.88 → 0.91**	**0.81 → 0.83**
	HPAP	36	**0.91 → 0.92**	**0.43 → 0.46**	**0.83 → 0.84**	0.77 → 0.75	**0.90 → 0.91**	**0.77 → 0.78**
	MESA	100	0.95	**0.46 → 0.59**	**0.85 → 0.87**	0.72 → 0.70	**0.89 → 0.91**	**0.77 → 0.80**
	MROS	134	**0.95 → 0.96**	0.44	**0.86 → 0.87**	**0.69 → 0.72**	**0.87 → 0.89**	**0.76 → 0.78**
	PHYS	100	0.84	0.60 → 0.57	**0.84 → 0.85**	0.81 → 0.79	**0.87 → 0.88**	0.79
	SEDF-SC	23	0.93	0.57	0.86	**0.71 → 0.76**	**0.87 → 0.88**	**0.79 → 0.80**
	SEDF-ST	8	**0.80 → 0.81**	**0.54 → 0.60**	**0.88 → 0.89**	**0.65 → 0.66**	0.91 → 0.90	**0.76 → 0.77**
	SHHS	140	0.94	**0.50 → 0.51**	0.87	0.77	0.91	0.80
	SOF	68	0.96	0.46 → 0.45	0.86	**0.76 → 0.77**	0.92	0.79
	WSC	218	0.89	0.53	0.90	0.61	0.90	0.77
	STAGES	89	0.82	0.37	0.81	0.69	0.81	0.70
	NCHSDB	102	0.89	0.38	0.86	0.90	0.83	0.77
Hold-Out	ISRUC-SG1	100	**0.88 → 0.90**	0.50	0.79 → 0.78	0.78 → 0.72	**0.89 → 0.90**	0.77 → 0.76
	ISRUC-SG2	16	**0.83 → 0.85**	0.50 → 0.49	0.78 → 0.76	0.82 → 0.74	**0.86 → 0.87**	0.76 → 0.74
	ISRUC-SG3	10	**0.87 → 0.90**	**0.56 → 0.57**	0.78 → 0.74	0.74 → 0.62	0.86	0.76 → 0.74
	MASS-C1	53	0.94	0.39 → 0.36	0.81	0.61	**0.88 → 0.90**	0.73 → 0.72
	MASS-C3	62	0.93	0.50	**0.85 → 0.86**	**0.73 → 0.74**	**0.91 → 0.92**	**0.78 → 0.79**
	SVUH	25	0.81	0.34 → 0.29	0.81	**0.82 → 0.86**	**0.88 → 0.89**	0.73
	DOD-H	25	0.92 → 0.90	**0.60 → 0.62**	**0.87 → 0.88**	**0.79 → 0.82**	0.94 → 0.93	**0.82 → 0.83**
	DOD-O	55	0.90	0.52 → 0.51	**0.86 → 0.89**	**0.74 → 0.78**	**0.92 → 0.93**	**0.79 → 0.80**
	DCSM-N	82	0.97	0.50 → 0.48	0.84	**0.78 → 0.80**	**0.87 → 0.88**	**0.79 → 0.80**
	DCSM-PLM	41	0.98	0.45	0.84	**0.75 → 0.76**	0.91 → 0.90	0.79
	DCSM-RBD	34	0.96	0.43 → 0.42	0.83 → 0.82	0.70	**0.74 → 0.81**	**0.73 → 0.74**
	DCSM-PD	24	**0.94 → 0.95**	0.37 → 0.35	0.70 → 0.67	0.62 → 0.60	**0.57 → 0.65**	**0.64 → 0.65**
	DCSM-RBD-PD	31	0.91 → 0.90	0.31 → 0.29	0.70 → 0.68	0.59 → 0.55	**0.43 → 0.57**	**0.59 → 0.60**
	Mean (weighted)		**0.93 → 0.94**	**0.50 → 0.51**	**0.84 → 0.85**	**0.76 → 0.77**	**0.88 → 0.89**	**0.78 → 0.79**
	STD (weighted)		0.04	0.07 → 0.08	0.04 → 0.05	0.07 → 0.08	**0.08 → 0.06**	0.04 → 0.05
	Per subject median		0.95	0.51	**0.86 → 0.87**	0.79	**0.91 → 0.92**	0.78
	Per subject MAD		0.03	0.11	0.05	0.11 → 0.12	0.04	0.06
	Pairs w. diff $\neq 0$, n		1504	1492	1513	1409	1444	1524
	Wilcoxon, W		398981	501496	459330	473064	346207	470123
	P-value		< 0.001	< 0.001	< 0.001	0.122	< 0.001	< 0.001

Table 11.2: Consensus score results on datasets (a) DOD-H and (b) DOD-O. The highest scores from human experts and the U-Sleep are highlighted in bold. Numbers shown are mean ± 1 standard deviation per-subject F1 scores computed between the output of a single model or human expert and the consensus scores generated from the 4 $(N-1)$ remaining (when comparing human to consensus) or best (when comparing the model to consensus) human annotators.

(a) DOD-H: Healthy controls, $N = 25$

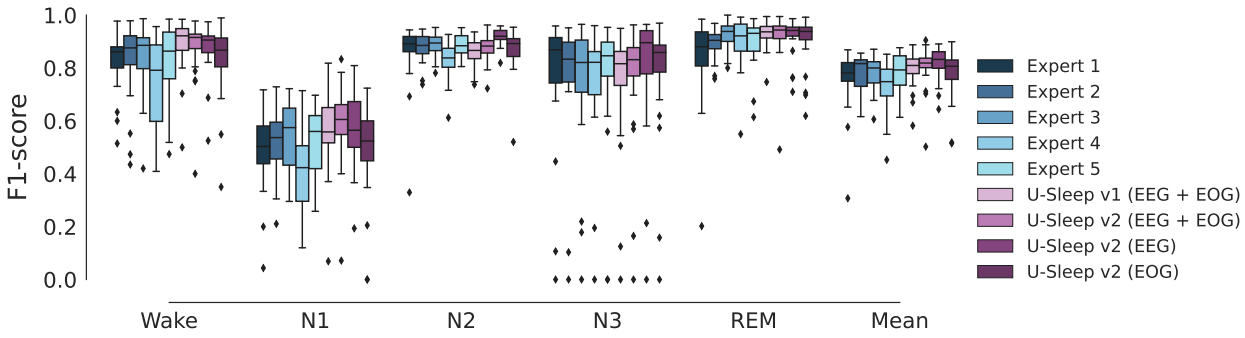
Scorer	Wake	N1	N2	N3	REM	Mean
Expert 1	0.83 ± 0.11	0.49 ± 0.15	0.86 ± 0.12	0.78 ± 0.24	0.84 ± 0.16	0.76 ± 0.11
Expert 2	0.83 ± 0.14	0.52 ± 0.11	0.88 ± 0.05	0.78 ± 0.23	0.89 ± 0.06	0.78 ± 0.07
Expert 3	0.84 ± 0.12	0.54 ± 0.13	0.88 ± 0.05	0.74 ± 0.25	$\mathbf{0.93 \pm 0.05}$	0.79 ± 0.07
Expert 4	0.73 ± 0.18	0.40 ± 0.15	0.83 ± 0.07	0.75 ± 0.22	0.90 ± 0.09	0.72 ± 0.11
Expert 5	0.83 ± 0.14	0.53 ± 0.12	0.89 ± 0.04	0.76 ± 0.24	0.90 ± 0.09	0.78 ± 0.08
U-Sleep v1 (EEG + EOG)	0.88 ± 0.10	0.56 ± 0.14	0.86 ± 0.05	0.73 ± 0.23	$\mathbf{0.93 \pm 0.05}$	0.79 ± 0.06
U-Sleep v2	0.88 ± 0.11	$\mathbf{0.59 \pm 0.14}$	0.88 ± 0.05	0.76 ± 0.22	0.92 ± 0.09	0.80 ± 0.08
U-Sleep v2 (EEG)	0.88 ± 0.09	0.57 ± 0.14	$\mathbf{0.92 \pm 0.03}$	$\mathbf{0.82 \pm 0.23}$	0.92 ± 0.07	$\mathbf{0.82 \pm 0.06}$
U-Sleep v2 (EOG)	0.83 ± 0.14	0.49 ± 0.19	0.87 ± 0.08	0.78 ± 0.23	0.91 ± 0.10	0.77 ± 0.10

(b) DOD-O: OSA patients, $N = 55$

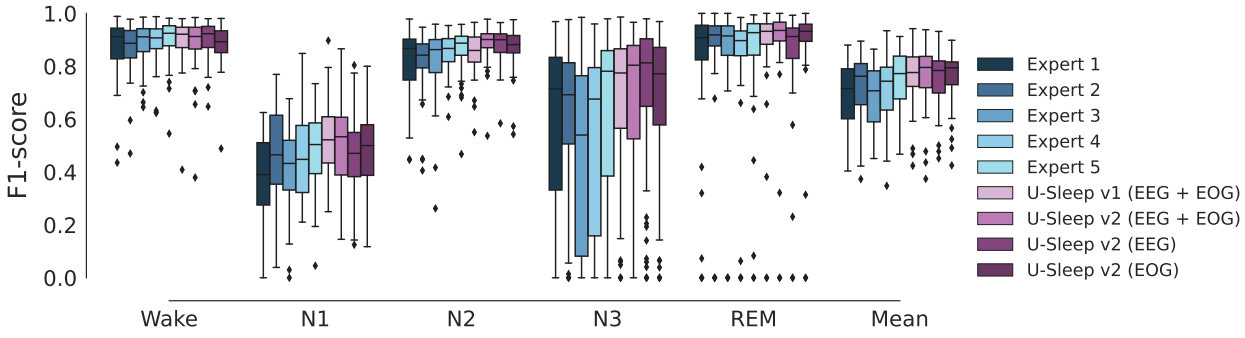
Scorer	Wake	N1	N2	N3	REM	Mean
Expert 1	0.87 ± 0.11	0.38 ± 0.15	0.82 ± 0.13	0.59 ± 0.31	0.81 ± 0.25	0.69 ± 0.12
Expert 2	0.87 ± 0.09	0.46 ± 0.17	0.82 ± 0.11	0.61 ± 0.29	0.86 ± 0.22	0.72 ± 0.12
Expert 3	0.88 ± 0.09	0.42 ± 0.16	0.83 ± 0.13	0.46 ± 0.33	0.85 ± 0.22	0.69 ± 0.11
Expert 4	0.89 ± 0.09	0.46 ± 0.15	0.84 ± 0.07	0.52 ± 0.33	0.83 ± 0.24	0.71 ± 0.12
Expert 5	$\mathbf{0.90 \pm 0.08}$	0.48 ± 0.15	0.86 ± 0.08	0.62 ± 0.33	0.85 ± 0.22	0.74 ± 0.11
U-Sleep v1 (EEG + EOG)	0.89 ± 0.09	$\mathbf{0.53 \pm 0.14}$	0.85 ± 0.08	0.66 ± 0.30	0.88 ± 0.20	$\mathbf{0.76 \pm 0.10}$
U-Sleep v2	0.89 ± 0.10	0.51 ± 0.16	$\mathbf{0.89 \pm 0.07}$	0.65 ± 0.31	$\mathbf{0.89 \pm 0.20}$	$\mathbf{0.76 \pm 0.11}$
U-Sleep v2 (EEG)	$\mathbf{0.90 \pm 0.07}$	0.47 ± 0.14	0.88 ± 0.07	$\mathbf{0.69 \pm 0.31}$	0.84 ± 0.20	$\mathbf{0.76 \pm 0.10}$
U-Sleep v2 (EOG)	0.89 ± 0.08	0.49 ± 0.15	0.87 ± 0.08	0.67 ± 0.30	0.88 ± 0.20	0.76 ± 0.10

Table 11.3: Mean ± 1 standard deviation sleep stage durations in hours computed from hypnograms scored by human experts and the U-Sleep v2 model at different staging frequencies on all $N = 2499$ PSGs of the validation- and test-set splits. Hypnograms were trimmed to include only stages between the first and last non-Wake scoring according to the scorings of the human expert. The Pearson's correlation r test/re-test value for sleep stage duration metrics are listed in parentheses, computed across a total $N = 473$ test/re-test comparisons of 282 unique subjects (all available multi-visit subjects). The highest correlation value(s) observed for each stage are highlighted in bold.

Annotator	Freq. (Hz)	Wake (hours)	N1 (hours)	N2 (hours)	N3 (hours)	REM (hours)
Human	1/30	1.41 ± 1.45 (0.43)	0.60 ± 0.50 (0.55)	3.43 ± 1.06 (0.45)	1.08 ± 0.88 (0.78)	1.18 ± 0.58 (**0.43**)
U-Sleep	1/30	1.42 ± 1.51 (0.45)	0.47 ± 0.38 (0.66)	3.54 ± 0.96 (0.48)	1.05 ± 0.77 (0.79)	1.22 ± 0.57 (0.37)
U-Sleep	1/6	1.52 ± 1.52 (**0.46**)	0.51 ± 0.37 (0.67)	3.37 ± 0.90 (0.50)	1.10 ± 0.75 (0.81)	1.20 ± 0.56 (0.37)
U-Sleep	1	1.54 ± 1.52 (**0.46**)	0.55 ± 0.37 (0.69)	3.21 ± 0.85 (0.51)	1.19 ± 0.74 (**0.83**)	1.20 ± 0.56 (0.37)
U-Sleep	5	1.56 ± 1.51 (**0.46**)	0.66 ± 0.39 (**0.70**)	3.01 ± 0.80 (**0.52**)	1.29 ± 0.74 (0.82)	1.18 ± 0.54 (0.37)



(a) DOD-H: Healthy controls, $N = 25$



(b) DOD-O: OSA patients, $N = 55$

Figure 11.1: Boxplots illustrating the distributions of F1 scores from 5 human experts and U-Sleep on healthy controls and OSA patients. Panel (a) shows results from dataset DOD-H on 25 healthy subjects. Panel (b) shows results from dataset DOD-O on 55 patients suffering from OSA. Sleep stages produced by U-Sleep and the five individual experts were compared to consensus-scored hypnograms. Please refer to the Methods section for further details. Mean F1 scores averaged across stages and F1 scores for the five individual sleep/wake stages are shown. The performance of different U-Sleep model versions is shown in shades of red (5 right-most boxplots in each group). The performance of each human expert is shown in shades of blue (5 left most boxplots in each group). Note that some records were scored by both human experts and U-Sleep with very low F1 scores (0 in some cases) on individual classes. This especially concerns stage N3 in dataset DOD-O and most often happens for rare classes. For instance, a patient severely affected by OSA rarely enters the N3 deep sleep stage. The resulting low number of observed N3 stages makes even a few errors resulting in a large deviation in the F1 score. Each boxplot shows the median (middle vertical line), first and third quartiles (lower and upper box limits) and whiskers that extend to 1.5 times the IQR added or subtracted from the third and first quartiles, respectively. Data outside of this range is marked as outliers indicated by diamond-shaped points.

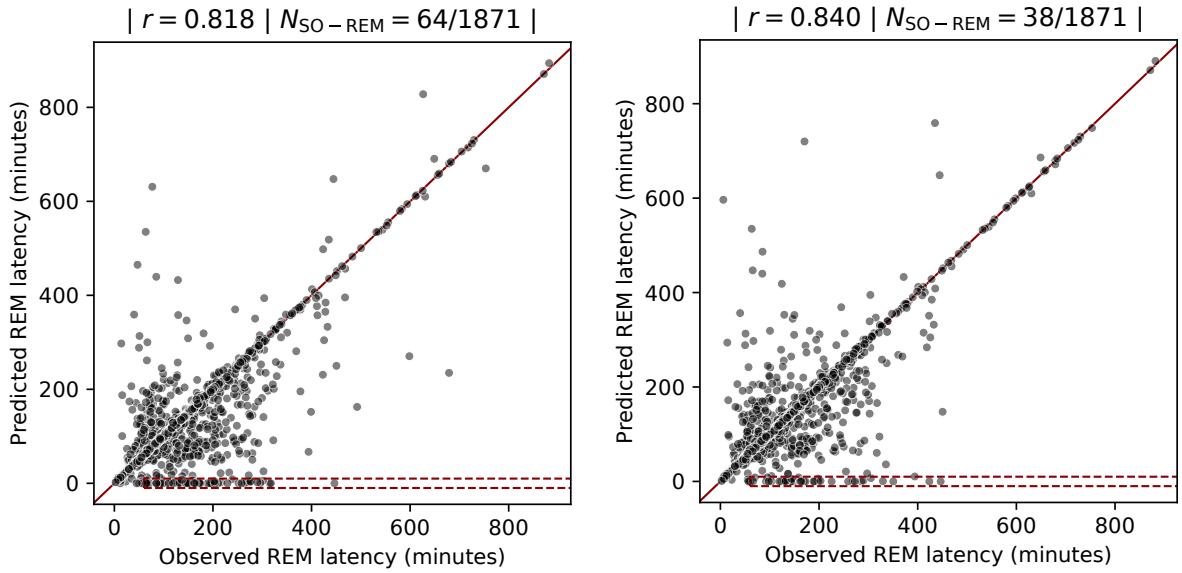
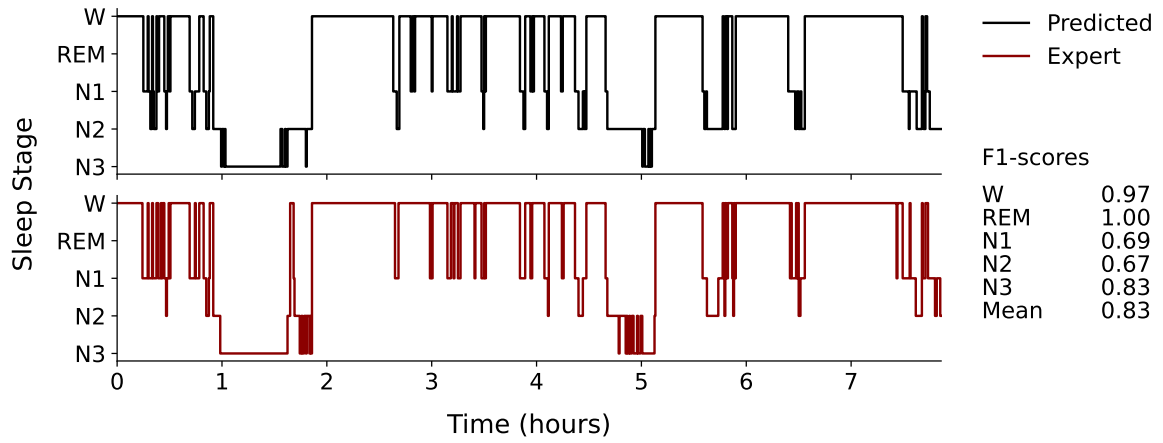
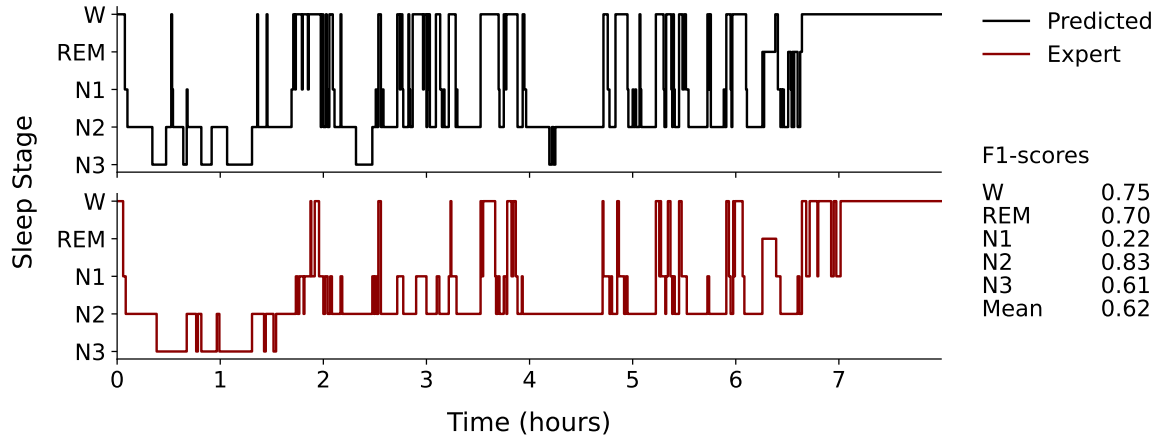


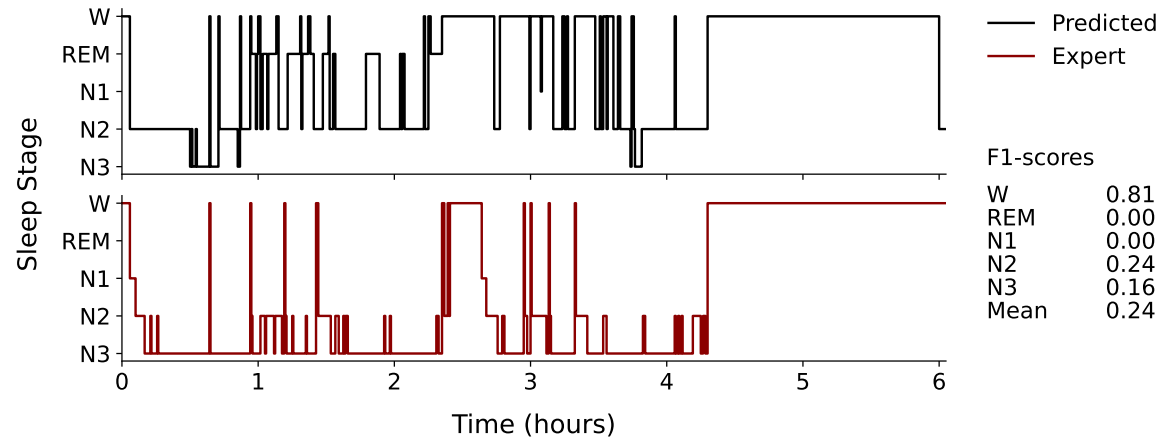
Figure 11.2: REM latencies computed from predicted majority-voted hypnograms of U-Sleep v1 (left) and U-Sleep v2 (right) correlated against observed REM latencies computed from expert annotated hypnograms. Pearson's correlation coefficient, r, is indicated above each plot, along with the number of likely wrongly predicted SO-REMs (dots within the lower red box of each plot; here defined by an observed REM latency of at least 60 minutes with a predicted latency of at most 10 minutes). The U-Sleep v2 performed REM latency estimation more accurately than the U-Sleep v1 with a higher general correlation and fewer wrongly predicted SO-REMs.



(a) Hypnogram with highest observed macro F1-score (record PD005).

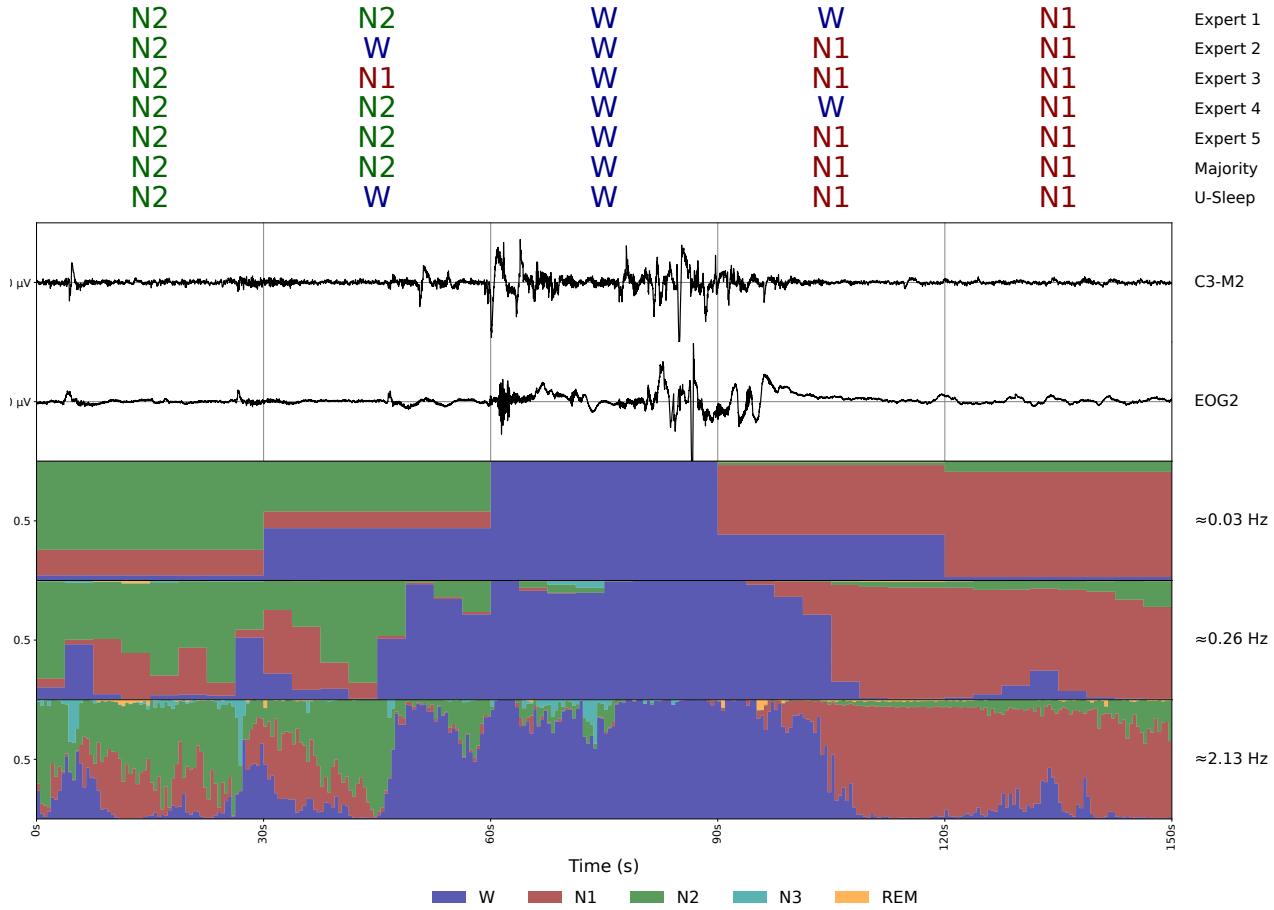


(b) Hypnogram with macro F1-score nearest dataset median (record PD033).

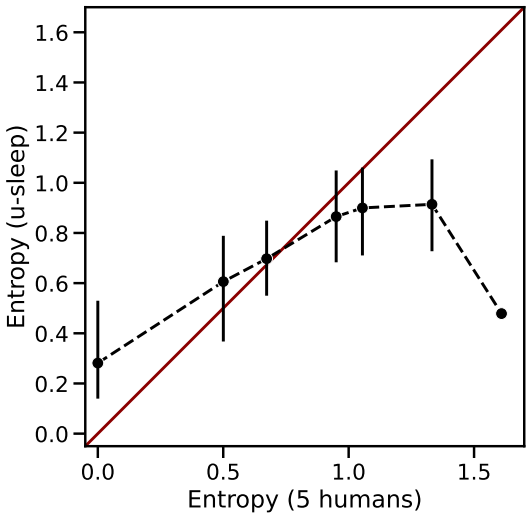


(c) Hypnogram with lowest observed macro F1-score (record PD018).

Figure 11.3: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across dataset DCSM-PD. Black hypnograms were predicted by U-Sleep, and red hypnograms are human expert annotations. F1 scores for each stage are shown to the right. Note that these unweighted per-subject F1 scores are noisy and may be misleading if the human annotator scores only a few instances of a given stage.

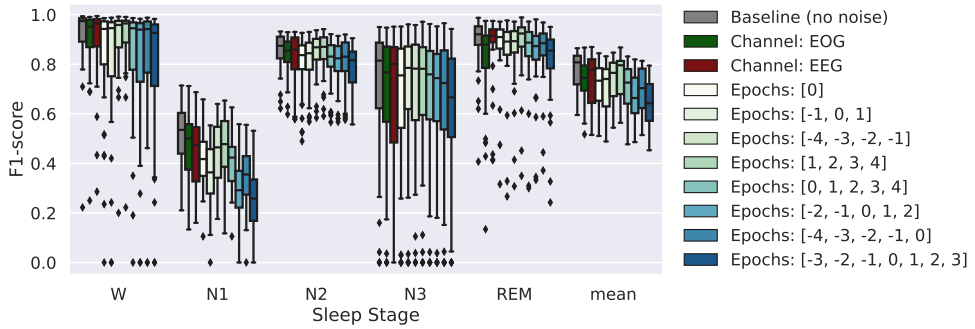


(a) Example, DOD-H subject. In epochs 2 and 4, the uncertainty of the group of human annotators was well captured by U-Sleep at 1/30 Hz scoring frequency. The possible source of human and model uncertainty was revealed at higher frequencies. U-Sleep separately scores the first and second half of each epoch 2 and 4 as different stages when not restricted to the arbitrary 30-second stage boundaries.

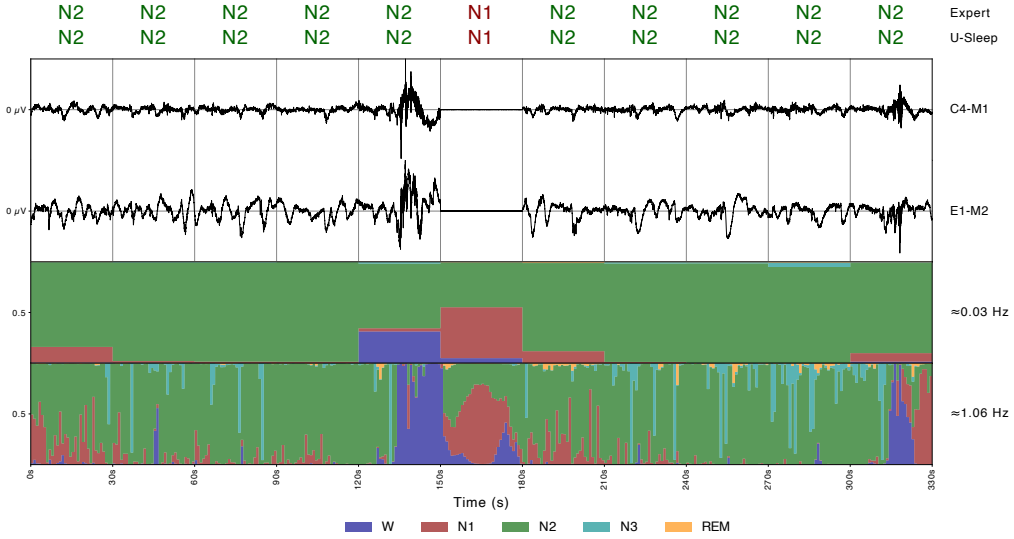


(b) Entropy experiment with U-Sleep v2 scores at 1/30 Hz.

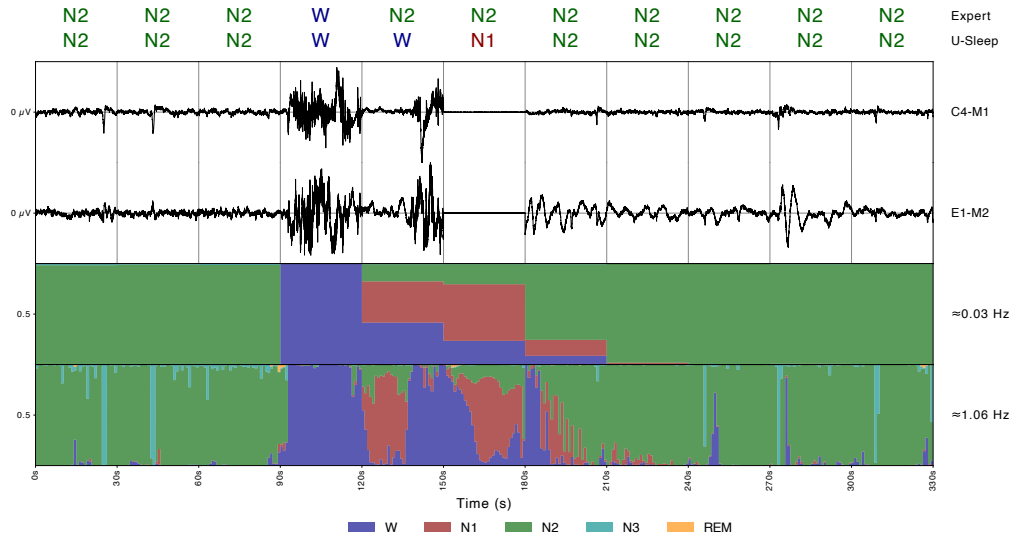
Figure 11.4: Entropy (uncertainty) correlation between U-Sleep v2 and a group of 5 human raters. $N = 80$, datasets DOD-O and DOD-H. See Methods for details on quantification.



(a) Boxplots showing F1 scores for each sleep stage computed across subjects under different context prediction experiments. The entire input modality was removed in the *Channel: EEG* and *Channel: EOG* experiments. In *Epochs* experiments, 30-second epochs were removed with indicated indices relative to the central-most epoch (the predicted). See Methods for details.

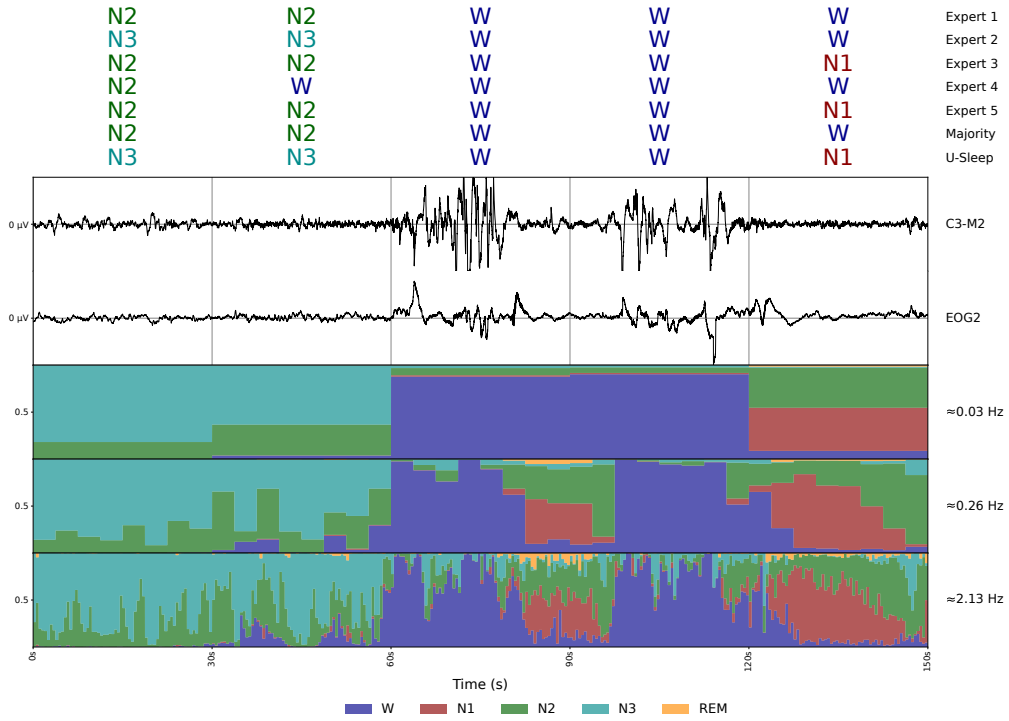


(b) Successful context prediction example where the central N1 stage was correctly guessed based on pre- or proceeding epoch data.

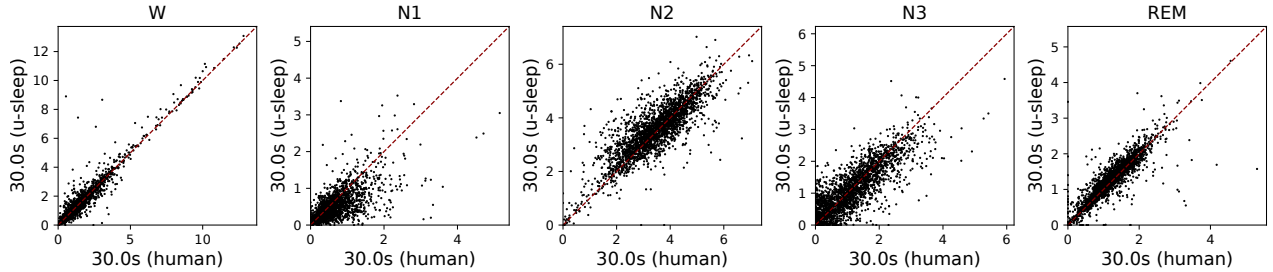


(c) Unsuccessful context prediction example where the central N2 stage was not correctly guessed based on pre- or proceeding epoch data. Missing information, the model assumed a smooth transition from Wake to N2 via an intermediate N1 stage.

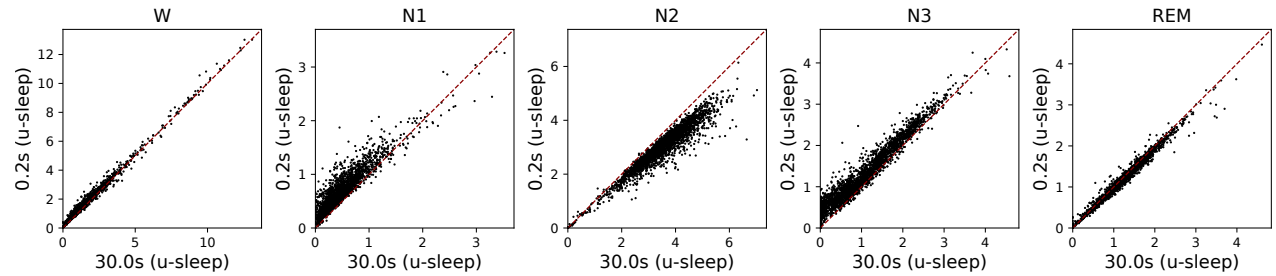
Figure 11.5: Context prediction experiment and illustrative examples for model U-Sleep v2 on $N = 39$ sleep studies from the `DCSM` dataset.



(a) Motivating example. Epochs 3 and 4 were scored as Wake by all expert annotators and U-Sleep at 1/30 Hz. U-Sleep detects a transient N1 or N2 stage at higher frequency outputs, a possible physiological phenomenon given the visible dampening in signal amplitude between epochs 3 and 4, which is lost in the compression of representing sleep in discrete 30-second blocks.

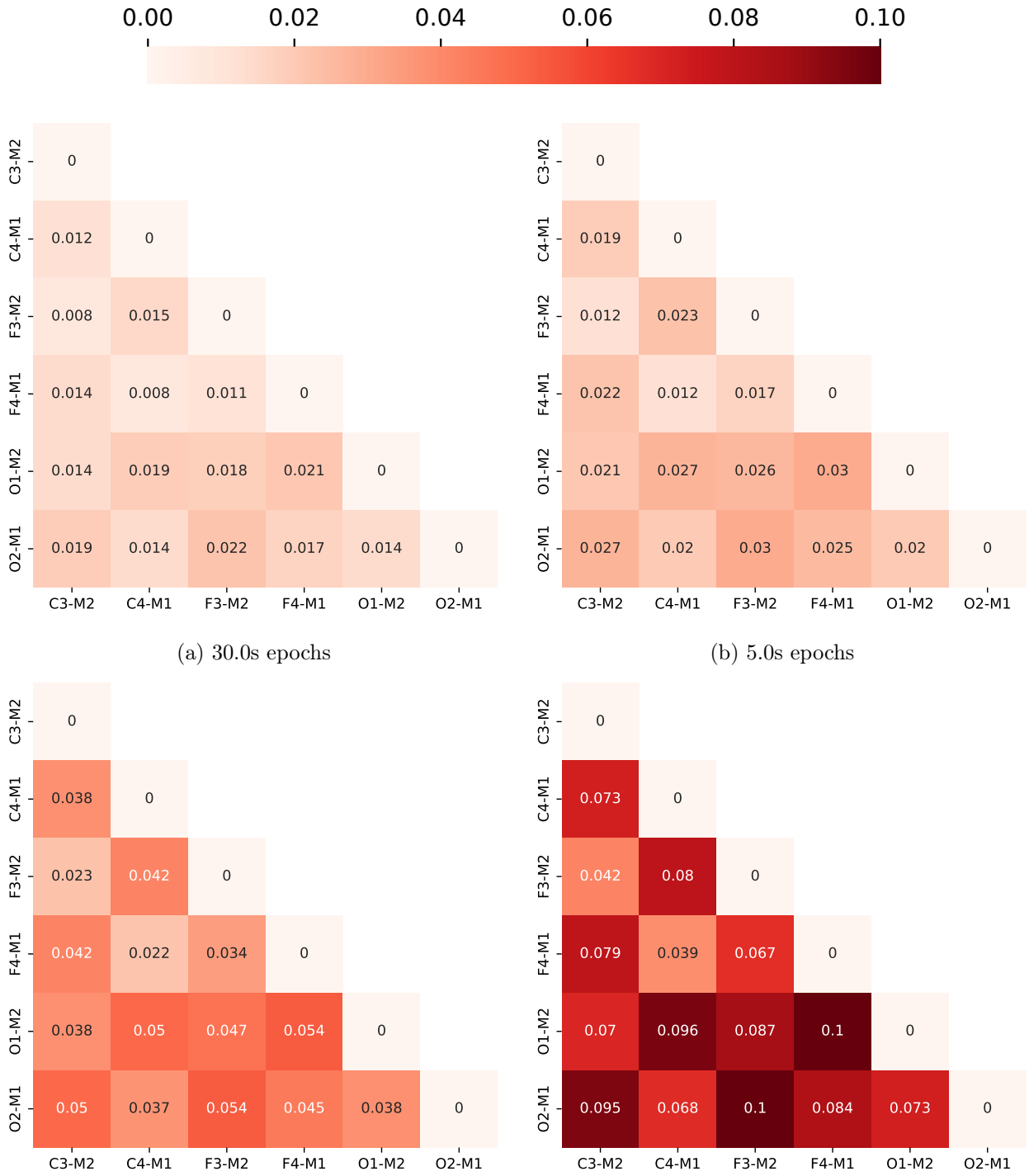


(b) Comparison of total sleep stage duration in hours measured by human experts at 1/30 Hz staging frequency and U-Sleep v2 at 1/30 Hz staging frequency on all $N = 2499$ PSGs from the validation- and test-set splits.



(c) Comparison of total sleep stage duration in hours as measured by U-Sleep v2 at 1/30 Hz staging frequency and U-Sleep v2 at 5 Hz staging frequency on all $N = 2499$ PSGs from the validation- and test-set splits.

Figure 11.6: Sleep stage duration as a function of staging frequency.



(a) 30.0s epochs

(b) 5.0s epochs

(c) 1.0s epochs

(d) 0.2s epochs

Figure 11.7: Spatial sleep scoring divergence (dis-similarity) experiment using the U-Sleep v2 model. Each number shows the mean epoch-wise Jensen-Shannon divergence (JSD) measure between the probability-like confidence score outputs of the model made in a pair of channels at different frequencies. Higher numbers (darker red) indicate that predictions were more diverging (i.e., less similar) on average. All $N = 842$ validation- and test-set PSGs with all channels available were used. Note that the colour bar is capped to the range [0, 0.10] to highlight differences, although the mean JSD measure in this experiment, in principle, could take a maximum value of $\log 2 \approx 0.693$ if orthogonal predictions were made in all epochs in all PSGs.

Chapter 12

Related work

The following pilot study on quantifying sleep stage confidence dynamics as a sleep stage transition is predicted by the U-Sleep v2 (EEG + EOG) model was initially planned for inclusion in Manuscript F, Chapter 11. It was removed because the methodological basis for how one should correctly and in a non-biased manner quantify the observed sleep stage transition speeds behaviour was not fully developed as of writing this thesis. See below for further discussions of the current approach's limitations. The section is instead presented here as an experimental pilot study of related work to inspire hopefully fruitful future research in the exciting domain of high-frequency sleep patterns.

12.1 Pilot study: Sleep stage transition speeds

12.1.1 Motivation

Visual examinations of high-frequency sleep stage scores revealed that the confidence of the U-Sleep model in a specific sleep stage might rise at varying *speeds* (i.e., over varying periods of time) whenever a stage transition occurs. See, for instance, Figure 12.1b and Figure 12.1a for examples of a *fast* and a *slow* stage transition, respectively. These dynamics will be referred to as *transition speeds*. While fast and slow transition speeds are difficult to relate directly to sleep physiology, as they may reflect the model's confidence in scoring difficult and easy stage transitions, it was hypothesised that the transition speeds might vary systematically and contain clinically relevant information. Consequently, an experiment was performed to quantify transition speeds and investigate differences between pairs of bi-directional stage transitions. In addition, transition speeds were correlated to demographic variables age, BMI and sex.

12.1.2 Methods

Stage transition speeds were computed for all $N = 2,499$ PSGs from the validation- and test-set splits of the dataset of manuscript F (chapter 11) using U-Sleep v2. Probability-like sleep scores were extracted at 1/5,Hz for quantifying slow transitions (defined as > 10 seconds) and 5,Hz for quantifying rapid transitions (defined as ≤ 10 seconds).

Fast transitions were quantified between peaks of high stage confidence defined by median confi-

dence of ≥ 0.5 in sliding windows over 27 confidence scores (5.4 seconds) at 5,Hz frequency. Transition speed was calculated only for pairs of peaks (a source peak, p_s, and target peak p_t) with non-identical stages and where p_s ended within 10 seconds of the start of p_t.

To quantify transitions over longer time scales, peaks were also extracted from raw (i.e., no median filter) 1/5,Hz predictions with ≥ 40 seconds continuous confidence of ≥ 0.35. Pairs of peaks ≤ 100 seconds apart were considered for further quantification.

A sigmoid function, defined as $f(x) = \frac{L}{1+\exp(-k(x-x_0))} + b$, was fitted to the 0.2 Hz or 5 Hz scores from the start of p_s to the end of p_t. Here, $f(x)$ models the confidence score in the stage of target peak p_t, x represents the relative time within the transition window, and $p = [b, L, k, x_0]$ is a parameter vector controlling the lower and upper range of $f(x)$, the steepness of the logistic function, and the x-axis offset of the sigmoid midpoint, respectively.

Parameters were fitted using the `scipy.optimize.curve_fit` function (Virtanen et al. 2020) with the *dogbox* method and a maximum of 1000 optimization steps. The initial parameter vector was set to $p_0 = [\text{med}(\mathbf{y}_s), \text{med}(\mathbf{y}_t), 1, m]$ for parameters b, L, k, and x_0, respectively. Here, $\text{med}(\mathbf{y}_s)$ represents the median confidence value within the source peak p_s, $\text{med}(\mathbf{y}_t)$ denotes the median confidence value within the target peak p_t, and m is the middle time point between the end of p_s and the start of p_t. The parameter bounds were set as $(0.0, 1.0)$, $(0.0, 1.0)$, $(-\infty, \infty)$, and $(\min \mathbf{x}, \max \mathbf{x})$ for parameters b, L, k, and x_0, respectively, where $\mathbf{x}$ is the vector of time-points (independent values in the fit).

Whenever a fit met the termination conditions within the maximum number of iterations, it was considered for further analysis. For each successful fit, the transition speed was defined and calculated as $s = f^{-1}(0.9L + b) - f^{-1}(0.1L + b)$ seconds, where f^{-1} is the inverse of the fitted sigmoid. The transition speed, s, indicates the number of seconds it takes for the (scaled/normalized) target stage confidence to increase from 10 % to 90 %, according to the fitted model $f(x)$. Figure 12.2 provides examples of detected source s_s and target s_t peaks along with the fitted logistic models for a fast and slow transition.

Transition speeds for all eligible transitions across sleep studies were visualized as histograms. To highlight potential non-symmetric transition speed dynamics (i.e., when transitioning from stage s_s to stage s_t is faster than the reverse transition from s_t to s_s), histogram bin-height differences were also calculated for all bi-directional stage transitions and plotted. Lastly, Pearson's correlation coefficients (r) were computed to assess the relationship between stage transition speeds and demographic variables age and BMI.

12.1.3 Results

A total of 444,593 sigmoidal fits (see Methods) were successfully fitted to model the U-Sleep v2 model's confidence scores transitioning from detected *source* peaks to *target* peaks. Figure 12.3 shows histograms of log-transformed transition speeds between all pair-wise combinations of source and target stages. Figure 12.4 shows bin-height differences between transition speed histograms computed for a stage transition $s_1 \rightarrow s_2$ and the reverse transition $s_2 \rightarrow s_1$.

As depicted in Figure 12.3, all sleep stage transitions followed either uni- or bimodal log-normal distributions. Table 12.1 presents the median, interquartile range (IQR), and the percentage of transitions classified as fast (i.e., less than or equal to 10 seconds). Transitions from all stages { N1, N2, N3, REM } to Wake followed approximately unimodal distributions with low median values (ranging from 1.6 to 2.8 seconds) and IQRs (ranging from 3.1 to 4.4 seconds). As a result, approximately 95 % of all observed transitions into stage Wake from any source were classified as fast. In contrast, transitions from { Wake, N1, N2, N3 } to REM were more frequently observed to be slow (i.e., with transition speeds greater than 10 seconds), with median values between 12.9 seconds (Wake to REM) and 46.7 seconds (N3 to REM), and IQRs ranging from 15.1 to 101.1 seconds. Additionally, transitions from REM to N1 (median of 39.5 seconds, IQR 51.0 seconds) and N2 (median of 29.3 seconds, IQR of 87.7 seconds) were notably slower than transitions between any pair of non-REM stages, which all had median values of less than 13 seconds. The transition from REM to the N3 stage had insufficient data points for a reliable analysis.

Some sleep stage transitions, most notably Wake to N2 and N2 to REM, followed approximately bimodal distributions. The separate detection of fast and slow transitions may have biased the identification of such bimodal distributions. However, it remains unclear why only certain transitions were affected in that case. See a further discussion of limitations below.

Figure 12.4 shows that most pairs of sleep stages exhibited non-symmetric bi-directional transition speed dynamics. All bi-directional Wake transitions (i.e., Wake $\rightleftarrows$ { N1, N2, N3, REM }) displayed a simple relationship where transitioning into Wake from any of the four other stages was, on average, faster than the opposite transition from any of the four stages into Wake. A similar pattern was observed for the bi-directional transition N1 $\rightleftarrows$ N2, where transitioning from N2 to N1 was, on average, faster than the reverse transition N1 to N2. Similar dynamics were observed for the bi-directional transition N2 $\rightleftarrows$ REM, although with less certainty due to fewer observations.

The transition dynamics for the bi-directional transition N2 $\rightleftarrows$ N3 were more complex, as depicted by multiple peaks in Figure 12.4g. Transitions from N2 to N3 were more commonly observed within the 0.1 – 1.0 seconds, 7.5 – 20 seconds, and $>$ 100 seconds ranges, while the opposite transition from N3 to N2 was more commonly observed in the intermediate 1.0 – 7.5 seconds and 20 – 100 seconds

Table 12.1: Sleep stage transition statistics. Each table cell displays the median and IQR (first line) and percentage (second line) of transitions classified as fast, i.e., less than or equal to 10 seconds for all pair-wise transitions (see distributions in Figure 12.3).

		Target Stage				
		Wake	N1	N2	N3	REM
Source Stage	Wake	-	5.2s - 5.3s 92 %	7.6s - 36.0s 61 %	3.2s - 7.1s 94 %	12.9s - 15.1s 58 %
	N1	2.1s - 3.9s 96 %	-	8.3s - 12.3s 72 %	-	43.1s - 101.1s 29 %
	N2	2.1s - 3.1s 96 %	5.6s - 6.6s 92 %	-	0.9s - 3.8s 91 %	17.9s - 68.3s 46 %
	N3	1.6s - 3.4s 94 %	13.4s - 15.0s 58 %	1.0s - 3.5s 93 %	-	46.7s - 79.3s 24 %
	REM	2.8s - 4.4s 96 %	39.5s - 51.0s 10 %	29.3s - 87.7s 37 %	-	-

ranges. Bi-directional transitions N1 $\rightleftarrows$ REM also exhibited non-simple dynamics, with transitions from N1 to REM more frequently observed in the < 10 seconds and > 100 seconds ranges, and the reverse transition REM to N1 more commonly observed in the intermediate 10 – 100 seconds range. The remaining bi-directional transitions (N3 $\rightleftarrows$ REM and N1 $\rightleftarrows$ REM) had too few observations to conclude their dynamics.

Pearson's correlation coefficients between transition speeds and demographic variables age and BMI were small (< 0.15) for all pairs of stage transitions except N1 $\rightarrow$ REM ($r = 0.39$) and N3 $\rightarrow$ REM ($r = 0.45$). Linear regression analyses revealed negative slopes of $a = -1.3$ and -0.8 for the two transitions, respectively, indicating an average decrease in transition speed with age. However, these transitions were particularly rare (at about 0.36 % and 0.12 % prevalence), and considerable standard errors on the slope estimates may suggest a random or weak correlation. Similarly, there were no clear correlations between transition speeds and variables BMI or sex.

12.1.4 Discussion and limitations

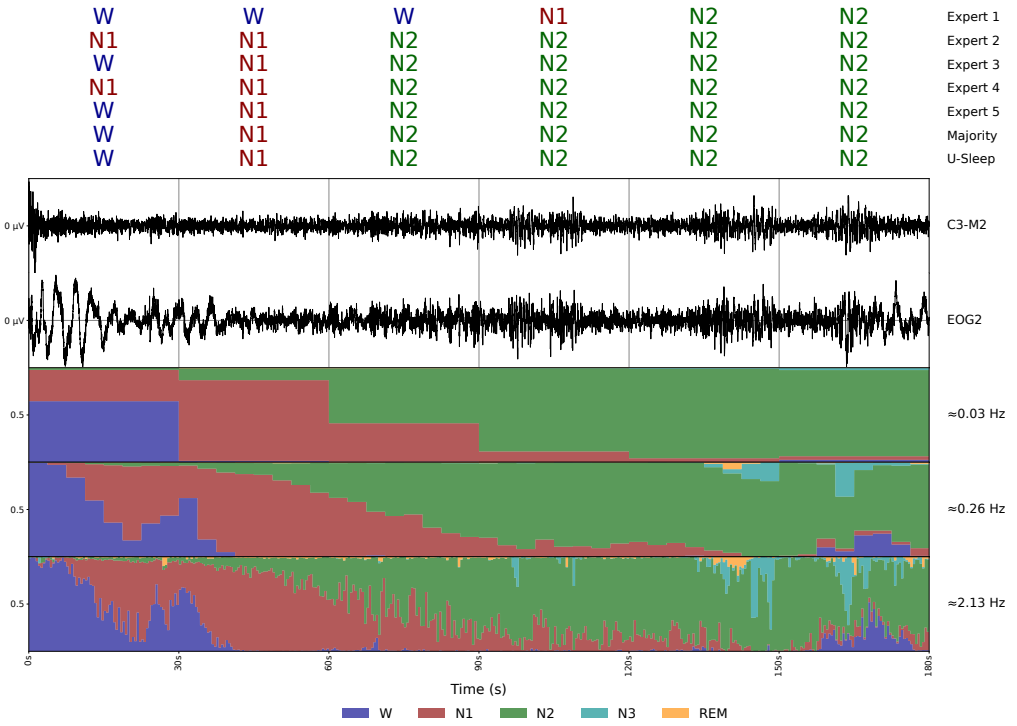
The analysis of sleep stage transition dynamics revealed several interesting patterns, such as general asymmetric transition dynamics where transitioning from one stage to another occurred rapidly in one direction and slowly in the opposite direction. However, no clear correlation was found between sleep stage transition speeds for any pair of stages and demographic variables age, BMI or sex. This raises questions about whether the calculated metrics truly reflect underlying physiology, as one might expect a genuine physiological phenomenon like sleep stage transition speeds to be influenced by variables like age, which is known to affect several aspects of sleep physiology.

These results may stem from the limitations and biases of the quantification method employed

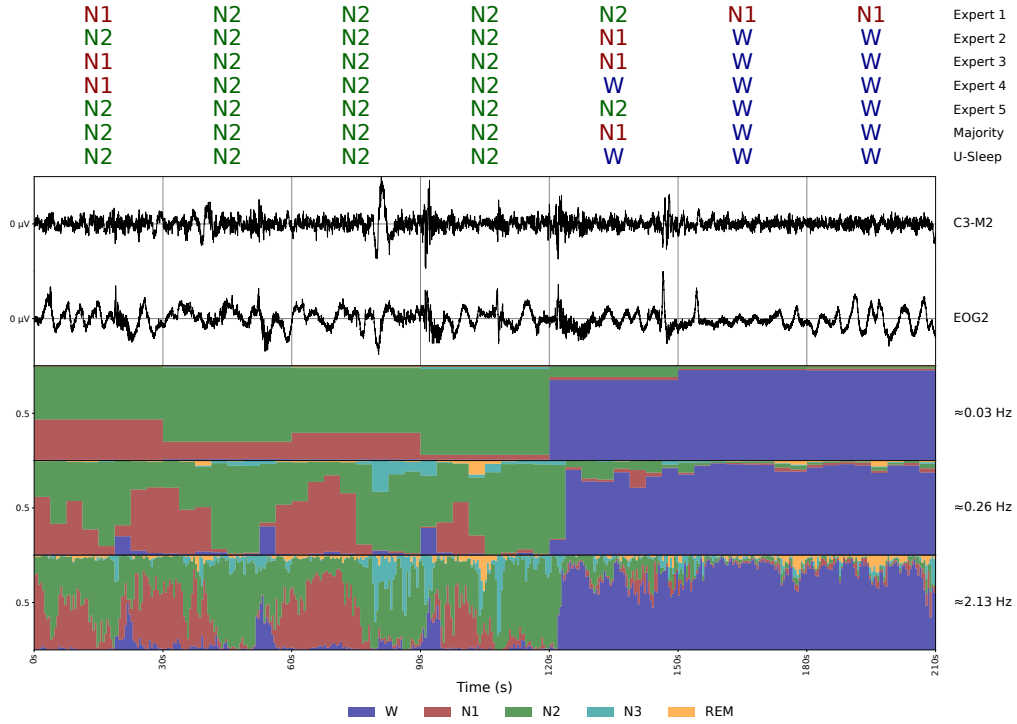
in this pilot study. First, the type and number of transitions detected were likely influenced by the chosen quantification method and its hyperparameters. The appropriateness of the selected sigmoidal model for the task is uncertain. Moreover, the hyperparameters were manually tuned based on a visual examination of randomly chosen fits to maximize the number of successful fits while minimizing false positive fits to non-relevant patterns. It was not feasible to manually investigate all the fitted sigmoidal models for correctness, and different parameters would likely have detected other types and numbers of transitions.

The impact of these biases on the results of the pilot study is unclear, such as whether the chosen quantification was biased towards detecting more fast or long transitions or if relevant intermediate transitions were missed. Nevertheless, the observation of non-symmetrical transition speeds between all pairs of stages remains noteworthy unless these potential biases affected the quantification of transition speeds systematically differently depending on the stage transition.

However, due to the limitations of the quantification method, this study was considered insufficiently mature for inclusion in Manuscript F, chapter 11. It is presented here to encourage future research in this area.

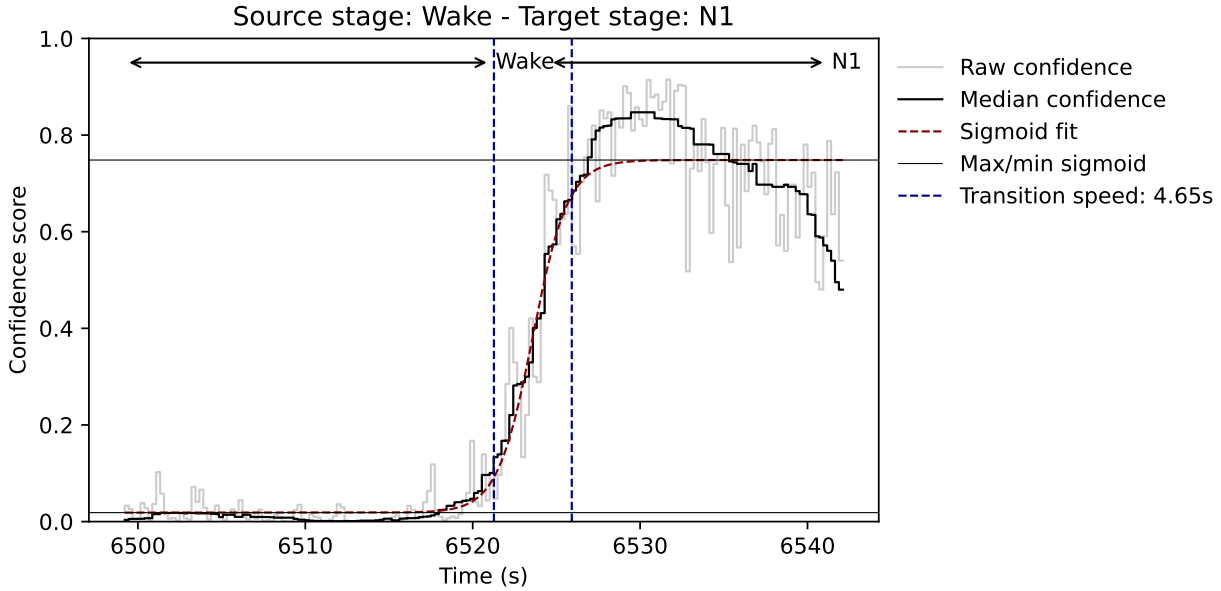


(a) Slow transition example. Between epochs 2 and 3, the transition N1 → N2 occurs. The confidence of the U-Sleep model increases in stage N2 from near 0 to near 1 over approximately 60 seconds.

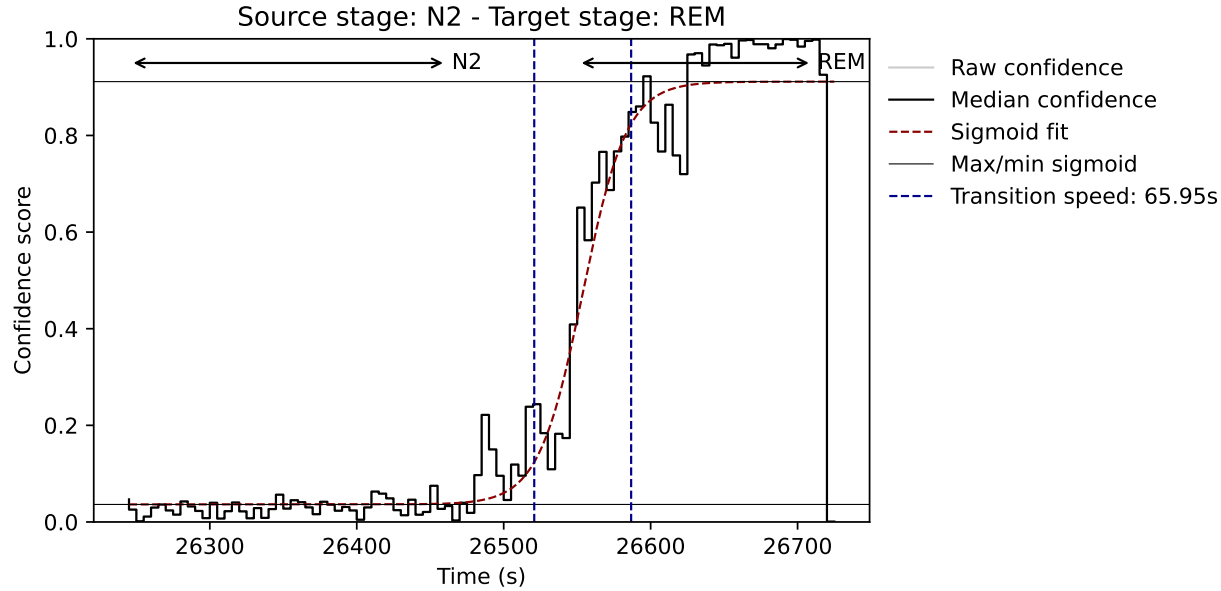


(b) Fast transition example. Between epochs 4 and 5, the transition N2 → Wake occurs. The confidence of the U-Sleep model increases rapidly in stage Wake from near 0 to near 1 over approximately 2-3 seconds.

Figure 12.1: Examples of slow and fast stage transitions as scored by the U-Sleep v2 model.



(a) Example of a fast transition from stage Wake to N1 quantified using a sigmoidal fit, giving a transition speed of ≈ 5 seconds.



(b) Example of a slow transition from stage N2 to REM quantified using a sigmoidal fit, giving a transition speed of ≈ 66 seconds.

Figure 12.2: Examples of fast and slow stage transitions quantified using a sigmoidal fit as described in the Methods section. Note that the plots in (a) and (b) have differently scaled x-axes.

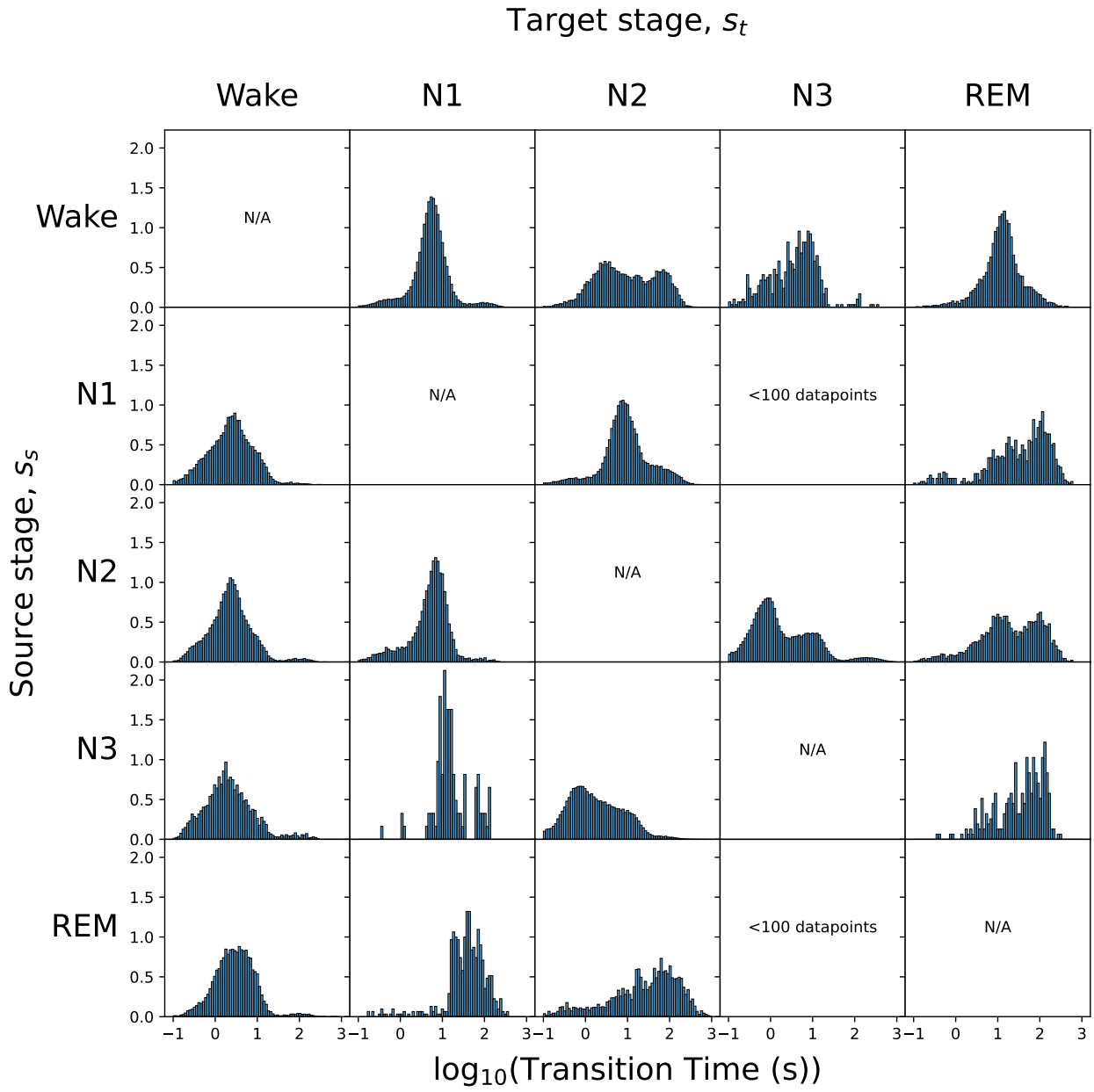
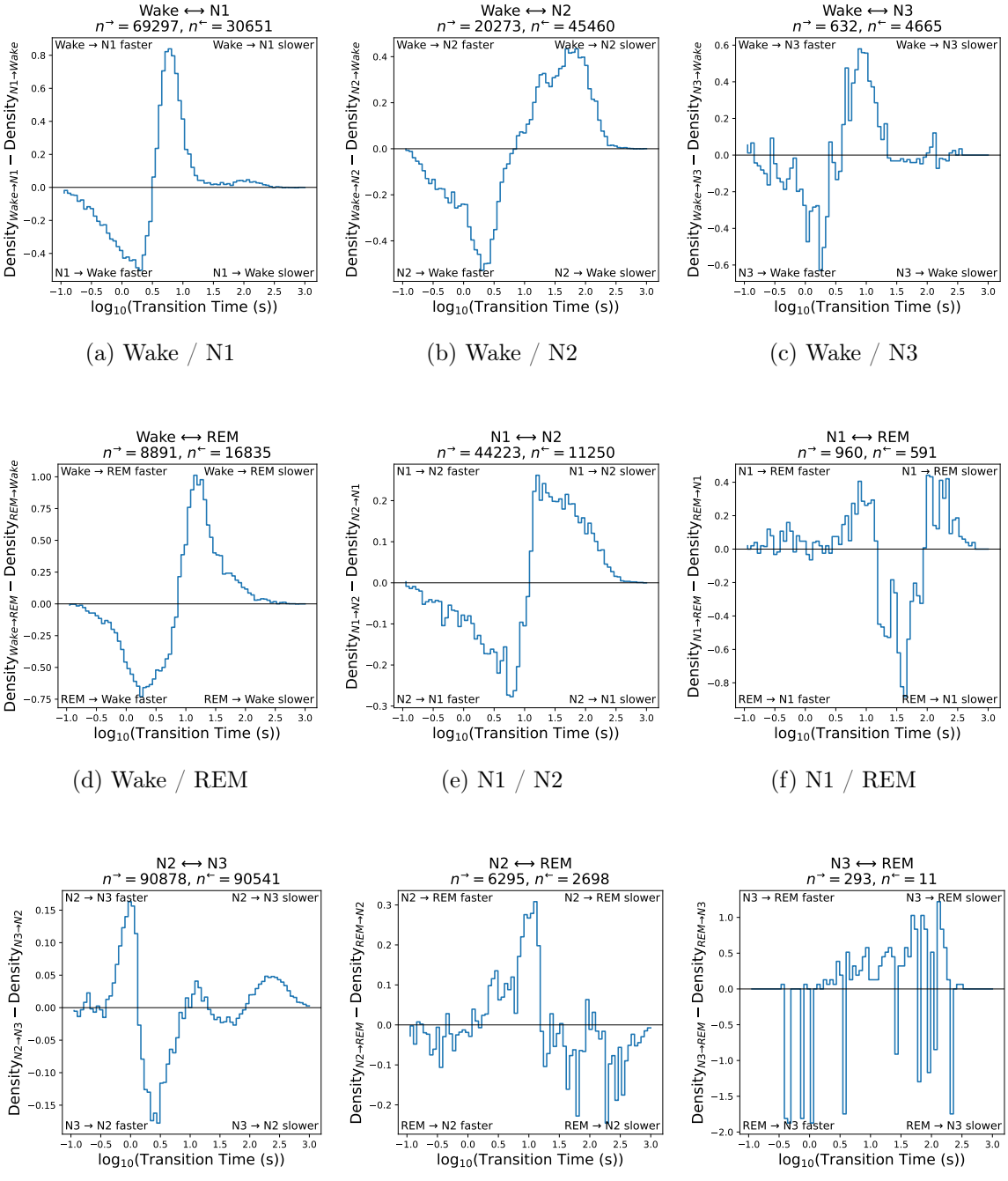


Figure 12.3: Pair-wise stage transition speeds. Each histogram shows $\log_{10}$-transformed transition times from a source stage (rows) to a target stage (columns). See Methods for details on the definition and quantification of transition speeds.



(a) Wake / N1 (b) Wake / N2 (c) Wake / N3

(d) Wake / REM (e) N1 / N2 (f) N1 / REM

(g) N2 / N3 (h) N2 / REM (i) N3 / REM

Figure 12.4: Transition speed histogram (see Figure 12.3) bin-height differences between bi-directional stage transitions. E.g., Figure 12.4a shows the difference between the transition speed histogram bins shown in Figure 12.3 for the transition Wake → N1 and reverse N1 → Wake. Each figure highlights the differences in average transition speed dynamics between any two stages.

Part IV

Discussion, Conclusions and Future Perspectives

Chapter 13

Discussion

13.1 Clinical robustness

Making clinically robust machine learning models for medical segmentation problems is challenging. Two primary difficulties stem from the complexity of medical data, which often varies with a patient's disease history and demographic, and may display systematic and sudden differences when new recording equipment or software is used to collect the data. The problem of high data variability can be minimized by training segmentation models on extensive and variable datasets. Collecting such datasets is, however, a non-trivial task. Even a model like U-Sleep, introduced in Paper D of Part III, trained on a massive and variable dataset collected by hundreds of researchers and medical doctors from several countries, cannot claim to have observed all relevant sleep data variability. It was, for instance, primarily trained on data from the EU and the US and data from a range of relevant patients, e.g., with severe brain disorders, were largely missing. Paper D, Abstract E and Manuscript F showed that U-Sleep could still generalize to new clinical sites and patient groups displaying complex EEG (e.g., PD patients) not represented in the training dataset, but this ability may be specific to the task of sleep staging and cannot be guaranteed in general.

Even if the training data included all imaginable clinical variability at a given instant, the model would likely deteriorate in performance with time because of continuous data drift and shift events. Physiological data is complex and may drift over time as demographic factors evolve or when new diseases emerge that change the characteristics of the recorded signals or images. Shift events, after which new data systematically differ from the original training data, are likely to occur as new hardware (e.g., new scanners or EEG recording equipment) or software updates are implemented over time.

For a concrete example of the problem of the inability to transfer machine learning models to other, even similar tasks in the medical domain, consider the single-cohort models developed in Paper B of this thesis on knee cartilage segmentation. Paper B did not intend to create model instances that are clinically robust across cohorts or MRI- scanners and sequences but instead studied the ability of the MPUNet pipeline to work reliably across such. However, following the paper's publication, researchers were interested in applying one of our models to their knee MRI data (direct communication; unpublished work). Their data was imaged on scanners and MRI sequences different from those considered in the paper. While the task was nearly identical and very narrow, i.e.,

segmentation of knee cartilages in MRI, the models of Paper B, which were found to perform at state-of-the-art accuracies in the single-cohort experiments, failed to produce meaningful results on the data of the external researchers. First, because of a technical difference in which the image volumes were differentially rotated and mirrored. Even after proper alignment, our developed models would produce nearly meaningless results, with most cartilage voxels wrongly predicted as background.

The intended application of the MPUNet would be to train a new model from scratch or fine-tune one of the models of Paper B on labelled data from the new dataset. However, the case exemplifies the barrier that end-users of machine learning models for healthcare tasks may encounter if they want to implement a pre-trained model developed for a similar but non-identical task. Unless very carefully communicated, end-users may not be aware that even minor differences between their data and the data with which the research model was trained may significantly decrease performance.

This issue affects all automatic segmentation systems. A set of hand-crafted rules in a classical system may no longer apply if data is collected from a new device or patient group. Typical machine learning systems cannot guarantee an expected performance if new data is not sampled from the same distribution that generated the training data. Therefore, one cannot simply develop and deploy a machine learning model in the clinic without considering how to generalise to clinical variability. Broadly speaking, one must adopt at least one of three (non-exclusive) techniques:

1. Re-training or fine-tuning the model to data specific to the clinical application site and the task. As studied in Part II, the machine learning pipeline must be easily applied without extensive hyperparameter re-tuning when transferred to new tasks or clinical sites to be applicable also when machine learning experts are unavailable. Because these methods demand labelled data for each new application, the pipeline should be as statistically efficient as possible.

   The fine-tuning approach is a simple transfer learning method that adapts a model to the target domain. The motivation is that the model should retain some of its learned information from the source domain and only need to adapt to select new features specific to the target domain. However, neural networks are prone to catastrophic forgetting, which may lead to overfitting to the small target domain dataset (French 1999), making it challenging to automate and generalize a transfer learning pipeline in clinical settings.

2. Attempt to map data collected in the target domain through various preprocessing steps to appear similar to data collected in the source domain. One may try to identify some quantifiable differences between data collected in the source- and target domains and create a pre-processing transformation pipeline which, for instance, attempts to standardize the intensity distributions of the images or attempts to establish a point-wise correspondence between anatomical landmarks (e.g., using image registration, see Oliveira et al. 2014 for a review).

3. Training the model on highly variable, cross-cohort datasets to induce model invariance properties to data variations which are unnecessary to solve the task, e.g., data variability introduced by different hardware equipment or software preprocessing steps.

These methods can – and probably should – be used in conjunction. For example, training on large and varied datasets, as done in Part III Paper D and Manuscript F, may create a clinically robust model at a specific instant in time. However, the model might not remain robust due to ongoing data drift and shift. As mentioned above, to address this, the model should be continuously updated by fine-tuning it on the original data and any new annotated data collected since its clinical implementation.

Arguably, any machine learning system to be clinically implemented should be evaluated concerning its cross-cohort performance and stability under data drift and shift events. Historically, most studies (including some of our own) did not perform such evaluations. It was common to consider only a single retrospective cohort collected from one clinical site and split it into parts for training and evaluation. While such papers have made significant technical contributions, few developed models have been clinically implemented as outlined in the Introduction Part I. Lately, however, increasing focus has been devoted to research on the training and evaluation of clinically robust machine learning models. The 2018 MSD challenge, for instance, was among the first to encourage the development of machine learning models that not only perform in a narrow domain (e.g., a single task on a single source of data) but also can solve many different tasks without requiring manual tuning from machine learning experts (Simpson et al. 2019). The work of this thesis emerged from the MSD. It investigated the development of clinically robust machine learning pipelines and instances of models for segmenting medical images and time series. Both approaches were feasible and will be discussed below.

13.2 MPUNet: A robust ML pipeline?

In Part II, the MPUNet, a pipeline that automatically trains a U-Net-like model for (near) arbitrary 3D medical image segmentation tasks using multi-planar data augmentation, was robust across most of the ten tasks of the MSD challenge as well as whole brain MRI, hippocampal MRI and knee cartilage MRI tasks. In contrast to most work at the time on automatically configuring pipelines, the MPUNet used a fixed model topology and hyperparameter set and did not rely on automatic hyperparameter experiments, AutoML or similar techniques (Y. He et al. 2021; Isensee et al. 2018). Further, the 2D U-Net, the backbone segmentation model of the MPUNet, although able to solve most segmentation problems when properly hyperparameter-tuned, does not guarantee sufficient clinical robustness alone. This can be seen, for instance, in Paper B of Part II on knee MRI segmentation,

in which the MPUNet compared favourably to a then state-of-the-art 2D U-Net when transferred without tuning to other clinical cohorts. The 2D U-Net trained without multi-planar augmentation transferred with lower performance.

These results indicate that the multi-planar augmentation scheme, the main differentiating feature of the MPUNet over the base U-Net, is the deciding mechanism that increases the clinical robustness of the MPUNet pipeline. As argued above, the base U-Net model can likely solve most segmentation tasks if its complexity and optimization are tuned correctly to the task and dataset size to ensure a proper fit (limited over- and underfitting). Empirically and based on qualitative investigations of learning curves, multi-planar training seems to reduce overfitting of even high-complexity U-Nets (e.g., with double the normal free parameters) on small datasets (e.g., 20-30 unique scans). This is likely because the mechanism drastically expands the availability of proper training data – i.e., samples from the actual training data rather than the typically deformed and potentially anatomically infeasible examples generated by other augmentations – while using a statistically efficient 2D model. Because the model operates in a lower-dimensional space than the training data from which numerous distinct examples may be generated, there are probably fewer functions in the hypothesis space that significantly overfits the more complex multi-view segmentation function.

However, this raises the question of why multi-planar training does not introduce significant underfitting. After all, the multi-planar target function must be more complex than any single-view function. We, however, argue that the multi-view function may not be considerably harder to approximate for three reasons: First, each additional view also introduces additional training data. If the views are significantly different, making the function harder to learn, the new training data is also more different, thus presenting a stronger learning signal. Secondly, the difficulty of learning new views does not scale linearly because of feature reusability. A significant fraction of the network will likely apply equally well to all views, e.g., convolution filters that extract low-level features such as edges and textures but also more complex ones such as shapes and some positional features (e.g., distance from the image centre), which are preserved across rotations. However, this is also likely why Paper A found the benefit of including additional views to eventually saturate, as each new view introduces less unique information. Third, some 3D structures may be easier to segment in particular views. For instance, structures with an easily recognizable shape when seen from a particular view. This benefit was further encouraged through the training of the fusion model, which weighs the contribution to the final ensemble prediction of each class from each view differentially, see Paper A. Finally, even if multi-planar training were to induce more errors when predicting in a single view, these may be cancelled out if uncorrelated across views due to the test-time ensembling that multi-view training allows.

In summary, enforcing equivariance to certain rotation transformations through multi-planar

augmentation seems to provide a beneficial inductive bias for segmenting 3D medical data using 2D FCNs, making the overall machine learning pipeline both computationally and statistically efficient and applicable for a wide range of tasks. Ultimately, however, a tradeoff must exist between under and overfitting. The above arguments are primarily intuitive and based on empirical observations, and the theoretical basis for why multi-planar training seems to work well in practice remains largely unresolved. For instance, while the term *rotational equivariance* was used loosely above, multi-planar augmentation only explicitly enforces equivariance to a small set of rotations depending on the chosen views and only for the base segmentation model. The separately trained fusion model oppositely breaks down the equivariance properties of the combined model (i.e., the base segmentation model and the fusion model) by applying a view-specific weighing of predictions made in each view for the final ensemble prediction. In Paper A, six views were empirically found to balance performance and (test-time) compute resources on the development dataset, and the views were randomly defined. However, optimal choices for how many views and which to use may depend on the task. It remains unclear how to efficiently choose views that maximize the accuracy of the ensemble prediction output by the base segmentation model and the fusion model in combination.

13.3 U-Sleep: A robust ML model?

The largely positive results of using FCNs with multi-planar training for medical images inspired our work on robust sleep staging in Part III by adapting the U-Net to 1D time series segmentation problems and later developing the U-Sleep model.

U-Time was adapted to 1D time-series segmentation problems with only minor modifications to the original U-Net for 2D image problems. Except for the apparent replacement of 2D convolutional operators with 1D counterparts, the number of filters in each convolutional layer was reduced, and just single convolution layers replaced the typical double-convolution encoder block to reduce the overall complexity of the model. The latter also because the stacking of multiple convolutional layers, each of which uses small kernel sizes such as 3×3, is often performed in image segmentation models primarily to exponentially increase the receptive field of the model with depth without introducing a squared expansion in the number of model parameters that increasing the size of a single 2D kernel would. This, however, is unnecessary in the 1D domain because the numbers of weights scale linearly with kernel size. Instead, the kernel sizes were increased in width, and so-called dilated convolutions were employed to expand the receptive field of the model. This was considered necessary for a task like sleep staging, where the input has long-range dependencies that may span several minutes of high-frequency (e.g., 128 Hz) signals.

High-frequency staging A so-called *segment classifier* module was also developed, which allows bridging the gap between the output of typical FCN models, which classify every position on the grid of the input data, and segmentation labels spanning more extended periods, which are characteristic of many time series tasks. The segment classifier averages over the high-frequency intermediate representation output by the encoder and decoder sub-networks. It then linearly combines the scores to produce a result for a period of interest. Another approach, which seems similar at first glance, would be to consider all data points belonging to the single label scored for a period of interest and train a typical FCN model against such dense or pseudo-high frequency label map. However, while not tested, we argue that it will introduce training instability and less valid high-frequency scores because the learning signal will discourage the detection of transient sleep stage changes or minority stages (i.e., stages that span less of the segment than another majority stage). In contrast, the segment classifier design does not restrict the model's ability to freely distribute class-confidence scores within the period of interest. Instead, it enforces that the average confidence over a given segment is the signal from which to produce the final score. Consequently, while not carefully studied here, the choice of aggregation function (i.e., the mean operation in the default U-Time implementation) may be essential and task-dependant, mainly if one is interested in studying the high-frequency segmentations that the U-Time architecture allows by changing the width of the aggregation function window to span a shorter period during prediction. The average function makes sense in sleep staging because it models the mental and AASM-established scoring guidelines of assigning the stage that spans the most time to a given segment. However, neither U-Time nor U-Sleep directly outputs the majority stage score but instead applies one (U-Time) or two (U-Sleep) linear layers on top of the average pooled scores to allow some flexibility. It remains, however, to be studied if these choices improved or decreased the overall scoring performance and how they affected the high-frequency outputs.

FCNs and LSTMs As hypothesised, the inductive biases of FCN models that make them suitable for image segmentation extended to time series segmentation problems like sleep staging. The feed-forward-only architectures were compute-efficient (processing a whole night's PSG data in a single forward pass at prediction time), easily optimized with limited overfitting, able to learn the long-range features necessary to solve the task effectively, and maintained high performance when transferred to new cohorts. In contrast, popular sequence models, such as mixed CNN and LSTM models, displayed decreased performance when transferred to other patient cohorts without hyperparameter tuning.

The increased stability and easier training of the feed-forward-only FCNs over recurrent architectures was hypothesised and unsurprising given the well-known difficulties of optimizing such models (Yoshua Bengio, Simard, et al. 1994; Pascanu et al. 2013). More surprisingly, U-Time was more

performant than mixed CNN and LSTM models on datasets where the recurrent models had been explicitly tuned to ensure a good fit. This indicates, in contrast to popular belief, e.g., in the automatic sleep staging community at the time, that explicit modelling of the temporal dynamics of sleep stages with recurrent units was necessary (or at least very useful) to learn proper sleep staging functions. Sufficiently deep FCNs can, however, also approximate arbitrarily complex temporal dynamics, which are explicitly encoded in the weights of their convolution kernels at deeper layers. Their special encoder-decoder architecture with skip connections allows such features to be computed at different temporal scales. Finally, FCNs naturally process long-range dependencies. U-Time, for instance, was numerically estimated to be responsive to input changes as far as 100.000 time-steps (approximately 13.5 minutes of 128 Hz signal) away from a given output (which could be easily extended using, e.g., larger or more dilated kernels). Such distant relationships are often challenging to model with recurrent models.

Cross-cohort & cross-channel training The U-Sleep model extended the U-Time model but also took inspiration from several other studies. The feasibility of simultaneous cross-cohort training was shown in our own Paper B on cross-cohort knee segmentation and pioneering studies such as Biswal, Sun, et al. (2018), which showed that machine learning-based sleep staging is feasible not only on extensive and heterogeneous cohorts but also that such models may be trained across multiple (two, precisely) cohorts simultaneously.

In addition, the special training of U-Sleep on randomly varying input EEG and EOG channel derivations has some resemblance to the multi-planar augmentation of Paper A. Instead of thinking of the sleep staging task as a mapping from a specific source of EEG activity to sleep stages, U-Sleep was, illustratively, trained in a multi-view fashion by instead considering the more general task of mapping a sleeping brain as *seen* from any EEG electrode derivation to sleep stages. Each EEG input can be considered an individual view of the physiological processes that inherently are the nature of the sleep stages we aim to score. Like the MPUNet, this training strategy further induces certain in- or equivariance properties, invariance under variable EEG and EOG input channels, specifically, which can be utilized to perform test-time augmentation to generate an average hypnogram of usually higher quality scored over multiple channel inputs. It should be noted that while multi-planar training encourages *equivariance* to view orientations, the cross-channel training of U-Sleep encourages *invariance* to channel derivations. In multi-planar training, the loss over the outputs of the model is minimized when the generated outputs are exact rotations of the ground truth label volume. In cross-channel training, the loss is minimized when the model produces predictions identical to the single ground truth hypnogram in all channels.

Clinical potential Based on Paper D and Manuscript F results, U-Sleep seems suitable for clinical sleep staging. It is robust to patient demographics, disease state and history, PSG recording equipment, EEG montaging, and data preprocessing. It was also found to perform at least at the human expert level on healthy and diseased patients from new clinical sites. It outperformed other automatic sleep staging models even if they were explicitly tuned and, more importantly, trained on data from those clinics. In Manuscript F, U-Sleep was further shown to produce reasonable sleep stage predictions for PSG recordings likely to display complex EEG patterns, such as from patients with RBD and PD. Abstract E showed that U-Sleep could even score sleep in acute stroke patients with good performance, considering how difficult human experts find this task, at least when grouping sleep into coarser Wake, REM and Non-REM stage groups. Manuscript F also found that the uncertainty of U-Sleep's outputs correlates with that of a group of five human experts, making these scores a useful study in clinical practice to direct the attention of human scorers towards difficult and ambiguous epochs. In summary, U-Sleep can score reliably under a wide range of clinical variability and likely serve as a useful tool for clinical sleep staging.

Single-channel staging & wearables The unintentional data loading, first discovered by Fiorillo, Monachino, et al. (2023) and described in the Supplementary Material F.1, introduced atypical channel derivations in the training of the original U-Sleep v1 model, causing interesting further robustness towards atypical input data. Specifically, these data induced additional invariance properties towards atypical EEG and EOG channel derivations. They also allowed using two EEG or two EOG inputs instead of the expected one EEG and one EOG combination. These results show that machine learning models like U-Sleep can solve sleep staging with robustness towards input data variability that would be difficult for human annotators to replicate and that machine learning models can score sleep based on robustly available features in near arbitrary EEG and EOG derivations. I.e., they do not require the typical, carefully defined PSG setup to function.

These findings inspired the development of single-channel U-Sleep variants, which can operate using only EEG or EOG inputs. While the dual-channel U-Sleep model performed slightly better, particularly on more complex cases, such as PD or RBD patients, the single-channel models were as accurate as even the best human rater out of five on healthy individuals and apnea patients. This indicates that U-Sleep may also support out-of-clinic and longitudinal studies through integration with, e.g., wearable devices. Several researchers are currently examining the use of U-Sleep in this domain and seeing promising early results, although these have not yet – to our knowledge – been published.

Novel sleep representations Manuscript F also showed that U-Sleep might be a useful research tool as the high-frequency and spatial sleep stage score output revealed interesting patterns and features. For instance, the outputs of U-Slee facilitated more robust sleep statistics (defined as more similar across multiple recordings of the same patient on different nights) compared to human experts (although not consistently; humans scored total REM sleep duration more robustly, see Manuscript F), and the robustness increased for most stages with higher staging frequencies. The special ability of U-Sleep to score in different EEG electrodes was found to reveal spatial sleep patterns at high-scoring frequencies, where scoring in spatially close electrodes on the same hemisphere was more similar than between distant & cross-hemispheric electrodes. However, it remains to be studied how and if these patterns provide additional clinical information. Coupled with the findings that U-Sleep's high-frequency outputs support the easier separation of patients from controls (sleep-disordered and acute stroke, Paper D and Abstract E), U-Sleep may be a valuable tool for studying sleep physiology and biomarkers in novel ways.

However, whether the high-frequency and spatial stages reflect underlying sleep physiology or indicate model-specific behaviour to different input signals remains uncertain. While Paper D, Abstract E, and Manuscript F provide some evidence in favour of the former, Part III's Related Work section discussed a pilot study on sleep stage transition speeds, which did not. This phenomenon, observed through high-frequency sleep stage scores generated by U-Sleep, seemed significant and potentially linked to a physiological phenomenon. However, we found no evident variation in these transition dynamics with any available demographic factors, such as age, BMI or sex, indicating that these specific dynamics may reflect model-specific uncertainties to different stages rather than physiology. However, the complex stage transition dynamics observed between some stages (e.g., N2 and N3) are not easily explained as only related to model uncertainty, and it is possible that the preliminary and rudimentary quantification of transition speeds limited our ability to detect correlations to demographic factors, should they exist.

Deployment U-Sleep was deployed to a free & unlimited (for research) use web server developed as part of this thesis at `https://sleep.ai.ku.dk/`. The hope was that it would facilitate global, large-scale studies of sleep, leveraging that automatic methods like U-Sleep are likely more consistent scorers than a pool of international human scorers. The service has been extensively used with more than 45.000 PSG scorings from more than 450 unique users. In particular, the recently developed API and its Python bindings (`https://github.com/perslev/U-Sleep-API-Python-Bindings`) have enabled a recent surge in large-scale studies conducted on multiple thousands of sleep studies. These include studies on specific patient groups, such as children and narcolepsy or Parkinson's patients, and the use of high-frequency staging to detect biomarkers (studies, to our knowledge, have

not yet been published).

With the deployment of U-Sleep as a free, high-performance and easy-to-use service, we also hoped to make expert knowledge on sleep staging available for people with limited access to well-trained medical doctors. However, the U-Sleep service is used primarily by researchers in Europe, North America and China, with large areas of the world, such as the majority of Africa and large areas of South America, seeing low numbers of users (stats based on anonymized analysis of data on users' affiliation to a research institution, which they submitted when creating their account). The usage may be proportional to the total number of PSGs conducted in different regions and shows, rather unsurprisingly, that while a tool such as U-Sleep may be made freely available, that alone does not ensure its adaptation if sleep recording facilities and practices are not already in place. Consequently, it is possible that U-Sleep would be more universally influential if implemented in easily accessible wearable devices that individuals at home can use.

13.4 Limitations

The studies presented in this thesis have several limitations, including both general and study-specific limitations:

- All models developed in this thesis were evaluated using only the single F1/Dice score metric for assessing the overlap between the automatically generated segmentation mask and a ground truth segmentation map. This metric is widely used in the machine learning community and is suitable for evaluating segmentations in particular because it ignores the often trivial true negative predictions (e.g., the trivial segmentation of the majority background in images). Being the harmonic mean of precision and recall, it further represents a conservative single-number aggregate that requires both to be high for the F1 score to be high.

  However, no single metric can fully capture the segmentation performance of a model, and the F1 score (or any other summary metric) may not truly reflect the clinical validity of the segmentation. Because all pixels of a particular class are weighted equally, the F1 score may be high even if a small but essential sub-region of the mask is poorly segmented. For instance, in knee cartilage thickness quantification, the F1 score may be high even if the thickness of the cartilage segmentation volume is not very accurate. In general, the F1 score favours large volumes, and errors near the boundary of a large segmentation volume may not be well reflected because correct predictions on the remaining volume outnumber errors on the boundary. The boundary is, however, of critical importance in many clinical scenarios as it may delineate, for instance, healthy- and tumour tissues or sleep stage transitions. For this reason, some

challenges, like the 2018 MSD (Simpson et al. 2019), also evaluated segmentation performance using an average surface distance metric.

Finally, the F1 can be highly affected by label imbalance. Other metrics exist that account for label imbalance. For instance, Coheen's Kappa, which is directly suitable for multi-class problems (Cohen 1960), or Matthews correlation coefficient (MCC), which can also be formulated for the multi-classification setting (Chicco et al. 2020; Gorodkin 2004). However, all metrics provide different views on the segmentation quality and make different assumptions.

In summary, one metric likely cannot stand alone when evaluating medical segmentations. The evaluations of this thesis could have been improved by evaluating multiple metrics and, more importantly, by studying the actual influence of the segmentation performance on clinical decisions based on them.

- More generally, this thesis did not fully evaluate the clinical relevance of the developed image- and sleep-stage segmentation methods, as the effects of using these systems in daily clinical routine were not studied. For instance, is the MPUNet feasible to apply for non-technical experts in the clinical setting, and how often does it require re-training to new data? How much does using U-Sleep decrease the time spent on manual sleep staging in practice? Several such questions would be interesting to investigate in future work; see also below.

- Models were generally evaluated concerning their average or median performance across subjects (with some exceptions, for instance, in Paper B, where the minimum scores are also assessed). However, in clinical practice, the worst-case performance may be equally important, at least if the system is fully automated or the operator cannot easily spot such outlier predictions. The possible adverse effect on the individual patient for which the system malfunctions should be considered before we can claim a system is fully clinically robust.

- As outlined in the Introduction part I, this thesis did not aim to address various other essential challenges for applying machine learning in healthcare, including ethical and legal barriers to implementation (who has the responsibility when the ML model fails?) and the lack of explainability of black-box methods (how did the model make its prediction for a given patient?).

- All evaluations were performed on retrospective data. It remains to be studied how, for instance, U-Sleep performs on prospective data that may be subject to gradual data drift and shift events (e.g., gradual cohort demographic changes and recording equipment upgrades).

- For the MPUNet pipeline specifically, it is an inherent limitation that the model instances that the channels produce are not themselves clinically robust and that the pipeline must be applied

to new, annotated data in each application domain. However, because of the shown ability of the MPUNet to learn variable tasks without hyperparameter-tuning, the pipeline can likely be used to develop robust model instances by training on large and heterogenous, cross-cohort datasets.

- For the U-Sleep model specifically, the study's limitation is that the model only scores sleep stages and not other sleep-related events. While U-Sleep solves the task of sleep staging with high accuracy and clinical robustness, it does not automate any of the remaining sleep-scoring tasks that humans perform when manually annotating PSG data in the clinical setting, such as the scoring of respiratory events, leg movements or cardiac events, which is necessary to diagnose certain disorders. Consequently, while U-Sleep automates a significant part of the PSG scoring process, manual inspection is still required whenever other non-sleep stage events are to be scored.

See the Future Perspectives chapter 15, which outlines possible research directions which may rectify some of these limitations.

Chapter 14

Conclusions

The overarching goal of this thesis was to develop clinically robust automatic segmentation models for medical images and time series. This problem was addressed in two different ways:

1. Part II studied clinically robust machine learning pipelines for medical image segmentation, i.e., a machine learning model, its optimisation procedure, and all hyperparameters controlling the two. To this end, we developed the MPUNet, a machine learning pipeline that trains an FCN under multi-planar data augmentation. The MPUNet was found clinically robust according to definition one of section 2.1 because it could be applied across many (although not all) tasks, clinical cohorts and scanner configurations with high performance without requiring extensive or manual hyperparameter tuning. This ability is due to the multi-planar augmentation scheme that efficiently uses the available 3D image information by re-sampling the proper training data distribution while training a statistically and computationally efficient 2D segmentation model.

2. Part III investigated clinically robust model instances for sleep staging. Our first finding was that FCN architectures are effective candidates for time series segmentation problems, often outperforming and exhibiting greater hyperparameter stability than mixed convolutional-recurrent networks. We then introduced the U-Sleep model, which was trained simultaneously on diverse sleep data from multiple cohorts and clinical sites. The model demonstrated clinical robustness per definition two in section 2.1, as it could be used without re-training on highly variable sleep data and achieved performance comparable to human expert annotators.

Concluding on the findings of Part II and III, the thesis made the following general observations on how to archive clinical robustness in deep learning segmentation models:

1. The general applicability of fully convolutional, feed-forward-only neural networks was reconfirmed. The U-Net base architecture performed well across diverse medical image segmentation tasks and was transferred to the new domain of time series segmentation with only minor modifications.

2. We found it beneficial to design data- augmentation and re-sampling techniques which induce model in- or equivariance properties to transformations of the input data that increase the clinical robustness of the model, even if the augmentations make the target function significantly

more complex, as long as the augmentations also considerably expand the set of actual training examples (i.e., samples from the proper training data distribution rather than deformed, augmented samples). The learned in- or equivariance properties can be used to perform test-time augmentation to output a more accurate ensemble prediction.

In Part II, multi-planar augmentation expanded the complexity of the target function, which should now map image volumes under multiple rotation transformations to a segmentation map, but also drastically expanded the available training data through re-sampling of the proper training data. The induced equivariance to (specific) rotations was further used to perform a single-model ensemble prediction on new data, eliminating segmentation errors that are decorrelated across views.

In Part III, training U-Sleep on randomly sampled input EEG and EOG channel derivations expanded the complexity of the target function, which should now map arbitrary input derivations to sleep stages, but also drastically expanded the available training data. The induced invariance to input derivations was further exploited by averaging multiple predictions on new data.

The MPUNet and U-Sleep were both trained using data augmentations which do not produce distorted examples but re-samples the available training data and use test-time augmentation to produce stronger segmentations on new data utilizing the induced in- or equivariance properties.

3. We found clinical robustness is achievable by training machine learning models on extensive and highly variable training datasets from several sources, even if each dataset varies concerning, for instance, recording hardware, patient population and data preprocessing pipeline.

This thesis also made the following software services and open-source software contributions:

1. An open-source (`MIT` licence) implementation of the MPUNet pipeline. It includes a command-line interface that allows model initialization, configuration, training and evaluation without code modifications. The software is available at `https://github.com/perslev/MultiPlanarUNet`

2. An open-source (`MIT` licence) implementation of the U-Time and U-Sleep models. It includes a command-line interface that allows model initialization, configuration, training and evaluation without code modifications. The software is available at `https://github.com/perslev/utime`.

3. A web service that deploys the U-Sleep pre-trained models for unlimited and free usage for any research application. The University of Copenhagen hosts the service at `https://sleep.ai.ku.dk/`. The service has a visual front-end as well as API for programmatic access. A set of

free Python bindings to the API were also made available at `https://github.com/perslev/U-Sleep-API-Python-Bindings` (`MIT` licence). As of April 2023, the service has been used to score more than 45,000 PSG files by more than 450 unique users, with an average of 250 predictions per day in 2023.

Chapter 15

Future perspectives

Medical image segmentation Several studies could be conducted to improve the MPUNet and our understanding of multi-planar augmentation:

- The MPUNet was developed alongside the widely used nnU-Net by Isensee et al. (2018) for the 2018 MSD MICCAI challenge (Simpson et al. 2019). Although the nnU-Net outperformed the MPUNet in the competition, the methodological advancements contributing to each system's clinical robustness are not mutually exclusive and could be combined. Multi-planar augmentation operates at the data level, while the automatic fingerprinting and model selection strategy of the nnU-Net function at the model architecture level. It would be worthwhile to explore the integration of multi-planar augmentation into the 2D candidate model of the nnU-Net. It would be interesting to study if this change increases the number of tasks for which the automatic model selection pipeline chooses the 2D (+ multi-planar augmentation) model candidate over the 3D and cascaded model candidates.

- As discussed in the Discussion section 13 above, it remains unclear how to efficiently estimate suitable views and the number of views for a given task. Currently, the MPUNet uses six randomly chosen views (with a restriction on minimum pairwise angles). A theoretical study of quantifying which views contain the most relevant information for a given task would likely improve the performance of the MPUNet. An empirical study of whether different segmentation tasks benefit from additional or fewer views and whether particular views, e.g., the canonical orthogonal image axes, should always be included in the set of views would also add valuable practical information.

- The base segmentation and fusion models are currently trained in two separate steps. This is because a whole segmentation volume must be predicted along each view and mapped back to the scanner coordinate space to establish point correspondence between the volumes for the fusion model to be trained. While a non-trivial task, if the combined model (the base model and the fusion model) could be cast into a single differentiable model and efficiently trained end-to-end, it may be possible to integrate the choice of view orientations into the learning problem itself so that suitable views that maximise the performance of the combined model are automatically learned. This would necessitate the design of a new type of fusion model that does not require point correspondence between the predictions within a single batch (likely

non-trivial) and a differentiable image interpolation module through which the loss can be backpropagated to update the set of current view vectors.

Sleep staging Several studies could be conducted to evaluate further and increase our understanding of the U-Sleep model for sleep staging and, in particular high-frequency and spatial sleep stages, and sleep medicine more broadly, including:

- A logical research direction is to extend U-Sleep to score other sleep-related events in addition to sleep stages. Several studies have addressed the automatic scoring of various sleep-related events in isolation; see, e.g., Nikkonen et al. 2021, Ferri, Zucconi, et al. 2005, Brink-Kjaer et al. 2020 and Acır et al. 2004 for automatic scoring of respiratory events, EEG sleep spindles, arousal and leg movements, respectively. Some studies have also developed methods that score multiple events simultaneously. For instance, Biswal, Sun, et al. 2018 implemented a mixed convolutional and recurrent neural network for scoring sleep stages, respiratory events, and limb movements. Several commercial systems can also score both sleep stages and other sleep-related events.

  A naive implementation to extend U-Sleep to score both sleep stages and other events would be to create an additional output from the model, which segments the PSG into short binary segments indicating the presence or absence of each event of interest. However, because many such events are sparse, non-contiguous and of variable length, the default FCN architecture may not be well suited for these tasks, as the dense binary segmentation of an arbitrarily defined grid may introduce heavy label imbalance, where the negative (no event) class is to be almost-always predicted. A better approach would be to use a model which can predict (any number of) event onset time points and associated durations. Transformer-based models may be suitable for such tasks and could be trained independently. However, due to the vast quantity of PSG data that has both sleep stages and other events annotated, better performance is likely achievable by training a single model, which can then leverage the additional learning signal from the expanded set of labels to learn features that support both predictive tasks. To benefit from the strengths of FCN architectures (for dense sleep stage segmentation) and Transformers (for sporadic event detection), the Transformer sub-network could be conditioned on the feature maps extracted by the encoder network of the FCN model. In reverse, the Transformer sub-network could affect the decoding of features into sleep stages in the FCN up-sampling sub-network, e.g., through an attention mechanism, which might also benefit the sleep scoring accuracy, as the occurrence of some events is highly indicative of a particular sleep stage (e.g., sleep spindles in stage N2).

- As described in Manuscript F, U-Sleep can operate on single input EEG channels (or EOG, with slightly lower average performance). An important future research route would be to evaluate and extend the U-Sleep model to various variable devices such as headbands or similar on which only a limited set of electrodes are available.

- Further work is needed to understand the relevance and potential of U-Sleep's high-frequency sleep stages. As shown in Paper D, Abstract E, Manuscript F, and the Related work section of Part III, there is likely clinically and physiologically relevant information in these scores, but the link between these scores and underlying physiology remains unclear. Future research could further study high-frequency sleeping patterns in health and disease by attempting to classify not just a single sleep disorder from a group of controls but in a more complex, realistic scenario in which several different sleeping disorders need individual classification. In addition, further research is needed to understand the implications of architectural modifications in the so-called segment classifier, e.g., the effect on high-frequency stages if aggregation functions other than the mean are used or how operations (such as learned linear combinations) applied on top of the aggregation output affects the final, high-frequency scores.

- Paper D and Manuscript F showed early hints at an interesting future research direction on studying spatial sleeping patterns using U-Sleep's ability to score sleep in EEG electrodes of arbitrary spatial position. In particular, it would be interesting to study whether distinct spatial sleeping patterns are robustly observable across individuals and whether they vary in health and disease as a function of staging frequency. These patterns could also be correlated to spatial observations of brain activity measured using fMRI in simultaneous EEG and fMRI studies (Mulert 2022).

- A modified version of U-Sleep could be developed to automate the task of sleep staging in other species, e.g., rodents.

15.1 Open data sharing initiatives

While the development and extensive clinical robustness of the U-Sleep model were made possible in part due to methodological advances, such as using FCNs for time series and training on randomly selected input channel derivations, what critically enabled this work was recent large-scale and open data sharing initiatives in the sleep data domain, including not least the National Sleep Research Resource (`https://sleepdata.org/`) and Physionet (`https://physionet.org/`) (Goldberger et al. 2000; G.-Q. Zhang et al. 2018). These resources collect not only vast quantities of labelled data, e.g., for training sleep stage classifiers, but also diverse data from many individual clinical

sites and patient cohorts while standardizing the technical specifications of the data to enable easier development of cross-cohort ML systems like U-Sleep. U-Sleep has shown the significant potential of such databases. We argue that clinically robust ML segmentation models may be created in a wide range of medical sub-fields and for a wide range of tasks if the already available annotated image and time series data are made available (in anonymized form and following all required ethical and legal approval processes) in centralized registers like the NSRR or Physionet. This potential was also shown in Paper 6 for segmenting knee cartilages in MRI, although in a small-scale experiment, where the MPUNet could be trained in a cross-cohort setup while retaining full performance on each cohort. Collecting significant numbers of variable MRIs from different clinical sites, scanners, MRI sequences, and patient cohorts may enable the development of robust cartilage segmentation models.

15.2 U-Sleep: Clinical certification and implementation

The U-Sleep model seems viable for clinical sleep staging because it is as accurate as human experts on PSG data from healthy and sleep-disordered individuals while coping with significant variability in patient demographics, recording equipment, EEG montages and preprocessing. U-Sleep has been implemented locally for research purposes in the Danish Center for Sleep Medicine at Rigshospitalet and received positive, qualitative feedback from expert scorers testing the system on a limited number of cases. There is an interest in verifying U-Sleep clinically and launching pilot studies of its applicability in daily clinical workflows and wearable devices.

However, several items must be addressed before U-Sleep can be made available to clinical end-users. First, U-Sleep needs clinical certification. For clinical usage within the European Economic Area (EEA), a Conformité Européenne (CE) mark must be issued for U-Sleep to be registered as a (likely) Class IIa medical device. The certification requires technical documentation of the safety and performance of U-Sleep as described in Annex I of the Regulation (EU) 2017/745 on medical devices (`https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0745`), which, among others, in chapter 1.1 states: *"They shall be safe and effective and shall not compromise the clinical condition or the safety of patients, or the safety and health of users or, where applicable, other persons, provided that any risks which may be associated with their use constitute acceptable risks when weighed against the benefits to the patient and are compatible with a high level of protection of health and safety, taking into account the generally acknowledged state of the art"*. The safety- and performance of U-Sleep must thus be evaluated to ensure the safety of the patients and users and weighed against the benefit that the device provides to clinical end-users and patients. Such risk analysis is submitted to a notified body and developed based on the results of a clinical investigation of the device. Concretely, U-Sleep must be studied in a prospective clinical study in which several

experiments may be relevant to ensure the safety and performance of the device:

1. The performance of U-Sleep should be assessed in multiple distinct clinical sites and on all relevant patient groups and demographics for which the system is ultimately intended. Such a study would be similar to those of Paper D and Manuscript F in Part III. Still, it must be conducted on new & prospective data using the deployment software implementation of the model in the relevant hospital environment.

2. The influence on the diagnostic process when using U-Sleep instead of manual scoring should be investigated to ensure that patients are not wrongly diagnosed when U-Sleep is implemented, e.g., in cases where U-Sleep fails to produce an accurate scoring. For instance, it may be relevant to study if the professional user's judgement is (negatively) affected by the availability of scores produced by U-Sleep.

3. Conversely, the time saved on manual sleep staging and potential benefits to diagnostic precision and consistency when using U-Sleep should be investigated to gauge the positive effects relative to any potentially discovered adverse effects.

U-Sleep may be clinically implemented if the benefit-risk ratio favours the patient and clinical end-users.

In late April 2023, BETA.HEALTH (`https://betahealth.dk/`) issued a grant of 500,000 DKK to support the clinical evaluation and implementation of U-Sleep. BETA.HEALTH is a joint Danish clinical innovation platform run by Rigshospitalet, Copenhagen, and Aarhus University Hospital, Aarhus, and sponsored by the Novo Nordisk Foundation (`https://novonordiskfonden.dk/en/`), The Central Denmark Region, and The Capital Region of Denmark. We greatly appreciate the support of BETA.HEALTH.

Bibliography

Ababneh, Saddam Y, Jeffrey W Prescott, and Metin N Gurcan. "Automatic graph-cut based segmentation of bones from knee magnetic resonance images for osteoarthritis research". In: *Med Image Anal* 15 (2011), pp. 438–448. DOI: 10.1016/j.media.2011.01.007.

Abadi, Martín et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015.

Abu-Mostafa, Yaser S., Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning From Data*. AMLBook, 2012.

Acır, Nurettin and Cüneyt Güzeliş. "Automatic recognition of sleep spindles in EEG by using artificial neural networks". In: *Expert Systems with Applications* 27.3 (2004), pp. 451–458. DOI: 10.1016/j.eswa.2004.05.007.

Agnew Jr, Harman W, James C Parker, Wilse B Webb, and Robert L Williams. "Amplitude measurement of the sleep electroencephalogram". In: *Electroencephalography and clinical neurophysiology* 22.1 (1967), pp. 84–86. DOI: 10.1016/0013-4694(67)90010-7.

Akkus, Zeynettin, Alfiia Galimzianova, Assaf Hoogi, Daniel L Rubin, and Bradley J Erickson. "Deep learning for brain MRI segmentation: state of the art and future directions". In: *Journal of digital imaging* 30 (2017), pp. 449–459. DOI: 10.1007/s10278-017-9983-4.

Albain, Kathy S, R Suzanne Swann, Valerie W Rusch, Andrew T Turrisi, Frances A Shepherd, Colum Smith, Yuhchyau Chen, Robert B Livingston, Richard H Feins, David R Gandara, et al. "Radiotherapy plus chemotherapy with or without surgical resection for stage III non-small-cell lung cancer: a phase III randomised controlled trial". In: *The Lancet* 374.9687 (2009), pp. 379–386. DOI: 10.1016/S0140-6736(09)60737-6.

Ambellan, Felix, Alexander Tack, Moritz Ehlke, and Stefan Zachow. "Automated Segmentation of Knee Bone and Cartilage combining Statistical Shape Knowledge and Convolutional Neural Networks: Data from the Osteoarthritis Initiative". In: 52.2 (2019). OAI-ZIB dataset, pp. 109–118. DOI: 10.12752/4.ATEZ.1.0.

Anderer, Peter et al. "Computer-assisted sleep classification according to the standard of the American Academy of sleep medicine: Validation study of the AASM version of the Somnolyzer 24 × 7". In: *Neuropsychobiology* 62.4 (2010), pp. 250–264. ISSN: 0302-282X. DOI: 10.1159/000320864.

Andreotti, F., H. Phan, N. Cooray, C. Lo, M. T. M. Hu, and M. De Vos. "Multichannel Sleep Stage Classification and Transfer Learning using Convolutional Neural Networks". In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2018, pp. 171–174. DOI: 10.1109/EMBC.2018.8512214.

Angelini, Elsa D, Olivier Clatz, Emmanuel Mandonnet, Ender Konukoglu, Laurent Capelle, and Hugues Duffau. "Glioma dynamics and computational models: a review of segmentation, registration, and in silico growth algorithms and their clinical applications". In: *Current Medical Imaging* 3.4 (2007), pp. 262–276. DOI: 10.2174/157340507782446241.

Anthony, Lasse F. Wolff, Benjamin Kanding, and Raghavendra Selvan. "Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models". In: *ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems*. 2020.

Arnal, Pierrick J. et al. "The Dreem Headband as an Alternative to Polysomnography for EEG Signal Acquisition and Sleep Staging". In: *bioRxiv* (2019). DOI: 10.1101/662734.

Assefa, Dawit, Harald Keller, Cynthia Ménard, Normand Laperriere, Ricardo J Ferrari, and Ivan Yeung. "Robust texture features for response monitoring of glioblastoma multiforme on-weighted and-FLAIR MR images: A preliminary investigation in terms of identification and segmentation". In: *Medical physics* 37.4 (2010), pp. 1722–1736. DOI: 10.1118/1.3357289.

Baandrup, Lone, Julie Christensen, Birgitte Fagerlund, and Poul Jennum. "Investigation of sleep spindle activity and morphology as predictors of neurocognitive functioning in medicated patients with schizophrenia". In: *Journal of Sleep Research* 28.1 (2018), e12672. DOI: 10.1111/jsr.12672.

Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun. "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling". In: *CoRR* abs/1803.01271 (2018). DOI: 10.48550/arXiv.1803.01271.

Bai, Shaojie, Zico Kolter, and Vladlen Koltun. "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling". In: *arXiv preprint arXiv:1803.01271* (2018).

Bakker, Jessie P., Ali Tavakkoli, Michael Rueschman, Wei Wang, Robert Andrews, Atul Malhotra, Robert L. Owens, Amit Anand, Katherine A. Dudley, and Sanjay R. Patel. "Gastric banding surgery versus continuous positive airway pressure for obstructive sleep apnea: A randomized controlled trial". In: *American Journal of Respiratory and Critical Care Medicine* 197.8 (2018), pp. 1080–1083. ISSN: 1535-4970. DOI: 10.1164/rccm.201708-1637LE.

Bates, Samuel, Trevor Hastie, and Ryan Tibshirani. "Cross-validation: What does it estimate and how well does it do it?" In: *arXiv* 2021 (2021), Preprint. DOI: 10.1080/01621459.2023.2197686.

Bauer, Stefan, Roland Wiest, Lutz-P. Nolte, and Mauricio Reyes. "A survey of MRI-based medical image analysis for brain tumor studies." In: *Physics in medicine and biology* 58.13 (2013). Place: England, R97–129. ISSN: 1361-6560 0031-9155. DOI: 10.1088/0031-9155/58/13/R97.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828. DOI: 10.1109/TPAMI.2013.50.

Bengio, Yoshua, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. "Greedy layer-wise training of deep networks". In: *Advances in neural information processing systems* 19 (2006).

Bengio, Yoshua, Patrice Simard, and Paolo Frasconi. "Learning long-term dependencies with gradient descent is difficult". In: *IEEE transactions on neural networks* 5.2 (1994), pp. 157–166. DOI: 10.1109/72.279181.

Berry, Richard B. and Mary H. Wagner. "Introduction". In: *Sleep Medicine Pearls (Third Edition)*. Ed. by Richard B. Berry and Mary H. Wagner. Third Edition. Philadelphia: W.B. Saunders, 2015, pp. 10–14. ISBN: 978-1-4557-7051-9. DOI: https://doi.org/10.1016/B978-1-4557-7051-9.00002-4.

Bilic, Patrick, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. "The liver tumor segmentation benchmark (lits)". In: *Medical Image Analysis* 84 (2023), p. 102680. DOI: 10.1016/j.media.2022.102680.

Bishop, Christopher et al. *Neural networks for pattern recognition*. Oxford university press, 1995.

Biswal, Siddharth, Joshua Kulas, Haoqi Sun, Balaji Goparaju, M. Brandon Westover, Matt T. Bianchi, and Jimeng Sun. *SLEEPNET: Automated Sleep Staging System via Deep Learning*. 2017.

Biswal, Siddharth, Haoqi Sun, Balaji Goparaju, M Brandon Westover, Jimeng Sun, and Matt T Bianchi. "Expert-level sleep scoring with deep neural networks". In: *Journal of the American Medical Informatics Association* 25.12 (Nov. 2018), pp. 1643–1650. ISSN: 1527-974X. DOI: 10.1093/jamia/ocy131.

Blackwell, Terri, Kristine Yaffe, Sonia Ancoli-Israel, Susan Redline, Kristine E. Ensrud, Marcia L. Stefanick, Alison Laffan, and Katie L. Stone. "Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: The osteoporotic fractures in men sleep study". In: *Journal of the American Geriatrics Society* 59.12 (2011), pp. 2217–2225. ISSN: 0002-8614. DOI: 10.1111/j.1532-5415.2011.03731.x.

Blum, Avrim and Tom Mitchell. "Combining labeled and unlabeled data with co-training". In: *Proceedings of the eleventh annual conference on Computational learning theory*. 1998, pp. 92–100. DOI: 10.1145/279943.279962.

Boccardi, Marina et al. "Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol". In: *Alzheimer's & Dementia* 11.2 (2015), pp. 175–183. DOI: 10.1016/j.jalz.2014.12.002.

Boostani, Reza, Foroozan Karimzadeh, and Mohammad Nami. "A comparative review on sleep stage classification methods in patients and healthy individuals". In: *Computer methods and programs in biomedicine* 140 (2017), pp. 77–91. DOI: 10.1016/j.cmpb.2016.12.004.

Bozinovski, Stevo. "Reminder of the first paper on transfer learning in neural networks, 1976". In: *Informatica* 44.3 (2020). DOI: 10.31449/inf.v44i3.2828.

Bragazzi, Nicola Luigi, Ottavia Guglielmi, and Sergio Garbarino. *SleepOMICS: How big data can revolutionize sleep science*. 2019. DOI: 10.3390/ijerph16020291.

Brandt, Martin et al. "An unexpectedly large count of trees in the western Sahara and Sahel". In: *Nature* 587 (2020), pp. 78–82.

Breiman, Leo. "Random forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

Brink-Kjaer, Andreas, Alexander Neergaard Olesen, Paul E. Peppard, Katie L. Stone, Poul Jennum, Emmanuel Mignot, and Helge B.D. Sorensen. "Automatic detection of cortical arousals in sleep and their contribution to daytime sleepiness". In: *Clinical Neurophysiology* 131.6 (2020), pp. 1187–1203. ISSN: 1388-2457. DOI: https://doi.org/10.1016/j.clinph.2020.02.027.

Buysse, Daniel J, Anne Germain, Martica Hall, Timothy H Monk, and Eric A Nofzinger. "A neurobiological model of insomnia". In: *Drug Discovery Today: Disease Models* 8.4 (2011), pp. 129–137. DOI: 10.1016/j.ddmod.2011.07.002.

Challen, Robert, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. "Artificial intelligence, bias and clinical safety". In: *BMJ Quality & Safety* 28.3 (2019), pp. 231–237. DOI: 10.1136/bmjqs-2018-008370.

Chambon, Stanislas, Mathieu N. Galtier, and Alexandre Gramfort. "Domain adaptation with optimal transport improves EEG sleep stage classifiers". In: *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. IEEE, 2018. ISBN: 9781538668597. DOI: 10.1109/PRNI.2018.8423957.

Chan, Martin, Tracy CH Wong, Aidan Weichard, Gillian Nixon, Lisa Walter, and Rosemary SC Horne. "Sleep macro-architecture and micro-architecture in children born preterm with sleep disordered breathing". In: *Pediatric Research* 87.4 (2020), pp. 703–710. DOI: 10.1038/s41390-019-0453-1.

Char, Danton S, Nigam H Shah, and David Magnus. "Implementing machine learning in health care—addressing ethical challenges". In: *The New England journal of medicine* 378.11 (2018), p. 981.

Chattu, Vijay, Md. Manzar, Soosanna Kumary, Deepa Burman, David Spence, and Seithikurippu Pandi-Perumal. "The Global Problem of Insufficient Sleep and Its Serious Public Health Implications". In: *Healthcare* 7.1 (2018), p. 1. ISSN: 2227-9032. DOI: 10.3390/healthcare7010001.

Chen, Qiming and Ren Wu. "CNN Is All You Need". In: *CoRR* abs/1712.09662 (2017).

Chen, Xiaoli, Rui Wang, Phyllis Zee, Pamela L. Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L. Jackson, Michelle A. Williams, and Susan Redline. "Racial/ethnic differences in sleep distur-

bances: The Multi-Ethnic Study of Atherosclerosis (MESA)". In: *Sleep* 38.6 (2015), pp. 877–88. ISSN: 1550-9109. DOI: 10.5665/sleep.4732.

Chicco, Davide and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". In: *BMC genomics* 21 (2020), pp. 1–13. DOI: 10.1186/s12864-019-6413-7.

Christ, Patrick Ferdinand, Mohamed Ezzeldin Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin D'Anastasi, et al. "Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2016, pp. 415–423. DOI: 10.1007/978-3-319-46723-8_48.

Çiçek, Özgün, Ahmed Abdulkadir, Soeren Lienkamp, Thomas Brox, and Olaf Ronneberger. "3D U-Net: learning dense volumetric segmentation from sparse annotation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer. 2016, pp. 424–432. DOI: 10.1007/978-3-319-46723-8_49.

Clevert, Djork-Arné, Thomas Unterthiner, and Sepp Hochreiter. "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)". In: *International Conference on Learning Representations (ICLR)*. 2016. DOI: 10.48550/arXiv.1511.07289.

Cohen, Jacob. "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46. DOI: 10.1177/001316446002000104.

— *Statistical power analysis for the behavioral sciences*. Academic press, 2013. DOI: 10.4324/9780203771587.

Cohn, Robert, GEORGE N RAINES, DONALD W MULDER, and META A NEUMANN. "Cerebral vascular lesions: electroencephalographic and neuropathologic correlations". In: *Archives of Neurology & Psychiatry* 60.2 (1948), pp. 165–181. DOI: 10.1001/archneurpsyc.1948.02310020061005.

Cootes, Timothy F., Gareth J. Edwards, and Christopher J. Taylor. "Active appearance models". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.6 (2001), pp. 681–685. DOI: 10.1109/34.927467.

Cootes, Timothy F., Christopher J. Taylor, David H. Cooper, and Jim Graham. "Active shape models-their training and application". In: *Computer vision and image understanding* 61.1 (1995), pp. 38–59. DOI: 10.1006/cviu.1995.1004.

Creutzfeldt, O-D, G Bodenstein, and JS Barlow. "Computerized EEG pattern classification by adaptive segmentation and probability density function classification. Clinical evaluation". In: *Elec-*

troencephalography and clinical neurophysiology 60.5 (1985), pp. 373–393. DOI: 10.1016/0013-4694(85)91012-0.

Cribari-Neto, Francisco and Achim Zeileis. "Beta regression in R". In: *Journal of Statistical Software* (2010). ISSN: 1548-7660. DOI: 10.18637/jss.v034.i02.

Crum, W. R., O. Camara, and D. L. G. Hill. "Generalized Overlap Measures for Evaluation and Validation in Medical Image Analysis". In: *IEEE Transactions on Medical Imaging* 25.11 (2006), pp. 1451–1461. DOI: 10.1109/TMI.2006.880587.

Cummings, Steven R. et al. "Appendicular Bone Density and Age Predict Hip Fracture in Women". In: *JAMA: The Journal of the American Medical Association* 263.5 (1990), pp. 665–668. ISSN: 1538-3598. DOI: 10.1001/jama.1990.03440050059033.

D'Amour, Alexander, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. "Underspecification presents challenges for credibility in modern machine learning". In: *Journal of Machine Learning Research* (2020). DOI: 10.48550/arXiv.2011.03395.

Dam, Erik B., Arjun Desai, Cem Deniz, Haresh Rajamohan, Ravinder Regatte, Claudia Iriondo, Valentina Pedoia, Sharmila Majumdar, Mathias Perslev, Christian Igel, et al. "Towards Automatic Cartilage Quantification in Clinical Trials–Continuing from the 2019 IWOAI Knee Segmentation Challenge". In: *Osteoarthritis Imaging* (2023), p. 100087. DOI: 10.1016/j.ostima.2023.100087.

Dam, Erik B., Jenny Folkesson, Paola Pettersen, and Claus Christiansen. "Automatic morphometric cartilage quantification in the medial tibial plateau from MRI for osteoarthritis grading". In: *Osteoarthritis Cartilage* 15 (2007), pp. 808–818. DOI: 10.1016/j.joca.2007.01.013.

Dam, Erik B., Martin Lillholm, Joselene Marques, and Mads Nielsen. "Automatic segmentation of high-and low-field knee MRIs using knee image quantification with data from the osteoarthritis initiative". In: *Journal of Medical Imaging* 2.2 (2015). ISSN: 2329-4302. DOI: 10.1117/1.JMI.2.2.024001.

Dam, Erik B., Jos Runhaar, Sita M Bierma-Zienstra, and Morten A Karsdal. "Cartilage cavity — An MRI marker of cartilage lesions in knee OA with data from CCBR, OAI, and PROOF". In: *Magn Reson Med* 80 (2018), pp. 1219–1232. DOI: 10.1002/mrm.27130.

Danker-Hopfe, Heidi, Peter Anderer, et al. "Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard". In: *Journal of Sleep Research* 18.1 (2009), pp. 74–84. ISSN: 09621105, 13652869. DOI: 10.1111/j.1365-2869.2008.00700.x.

Danker-Hopfe, Heidi, D. Kunz, et al. "Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders: IRR of sleep stage scoring in patients". In: *Journal of Sleep Research* 13.1 (2004), pp. 63–69. ISSN: 0962-1105. DOI: 10.1046/j.1365-2869.2003.00375.x.

de Bruijne, Marleen. "Machine learning approaches in medical image analysis: From detection to diagnosis". In: *Medical Image Analysis* 33 (2016). 20th anniversary of the Medical Image Analysis journal (MedIA), pp. 94–97. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2016.06.032.

De Gennaro, Luigi, Michele Ferrara, Giuseppe Curcio, and Riccardo Cristiani. "Antero-posterior EEG changes during the wakefulness–sleep transition". In: *Clinical neurophysiology* 112.10 (2001), pp. 1901–1911. DOI: 10.1016/S1388-2457(01)00649-6.

De Gennaro, Luigi, Michele Ferrara, Fabrizio Vecchio, Giuseppe Curcio, and Mario Bertini. "An electroencephalographic fingerprint of human sleep". In: *Neuroimage* 26.1 (2005), pp. 114–122. DOI: 10.1016/j.neuroimage.2005.01.020.

Deeley, MA, A Chen, R Datteri, JH Noble, AJ Cmelak, EF Donnelly, AW Malcolm, Luigi Moretti, J Jaboin, K Niermann, et al. "Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study". In: *Physics in Medicine & Biology* 56.14 (2011), p. 4557. DOI: 10.1088/0031-9155/56/14/021.

Desai, A. et al. "A multi-institute automated segmentation evaluation on a standard dataset: Findings from the international workshop on osteoarthritis imaging segmentation challenge". In: *Osteoarthritis and Cartilage* 28 (2020), S304–S305. DOI: 10.1016/j.joca.2020.02.477.

Desai, Arjun, Francesco Caliva, Claudia Iriondo, Aliasghar Mortazi, Sachin Jambawalikar, Ulas Bagci, Mathias Perslev, Christian Igel, Erik B. Dam, Sibaji Gaj, et al. "The international workshop on osteoarthritis imaging knee MRI segmentation challenge: a multi-institute evaluation and analysis framework on a standardized dataset". In: *Radiology: Artificial Intelligence* 3.3 (2021), e200078. DOI: 10.1148/ryai.2021200078.

Devore, Elizabeth E, Francine Grodstein, and Eva S Schernhammer. "Sleep Duration in Relation to Cognitive Function among Older Adults: A Systematic Review of Observational Studies". In: *Neuroepidemiology* 46.1 (2016), pp. 57–78. DOI: 10.1159/000442418.

Dice, Lee Raymond. "Measures of the Amount of Ecologic Association Between Species". In: *Ecology* 26.3 (1945), pp. 297–302. DOI: 10.2307/1932409.

Dong, Hao, Akara Supratak, Wei Pan, Chao Wu, Paul M. Matthews, and Yike Guo. "Mixed Neural Network Approach for Temporal Sleep Stage Classification". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26.2 (2018), pp. 324–333. DOI: 10.1109/TNSRE.2017.2733220.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020). DOI: 10.48550/arXiv.2010.11929.

Drinnan, Michael J., Alan Murray, Clive J. Griffiths, and G. John Gibson. "Interobserver Variability in Recognizing Arousal in Respiratory Sleep Disorders". In: *American Journal of Respiratory and Critical Care Medicine* 158.2 (1998), pp. 358–362. ISSN: 1073-449X, 1535-4970. DOI: 10.1164/ajrccm.158.2.9705035.

Drozdzal, Michal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. "The importance of skip connections in biomedical image segmentation". In: *International Workshop on Deep Learning in Medical Image Analysis, International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Springer. 2016, pp. 179–187. DOI: 10.1007/978-3-319-46976-8_19.

Eisenhauer, Elizabeth A, Patrick Therasse, Jan Bogaerts, Lawrence H Schwartz, Danielle Sargent, Robert Ford, Janet Dancey, S Arbuck, Steve Gwyther, Margaret Mooney, et al. "New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)". In: *European journal of cancer* 45.2 (2009), pp. 228–247. DOI: 10.1016/j.ejca.2008.10.026.

Elsken, Thomas, Jan Hendrik Metzen, and Frank Hutter. "Neural architecture search: A survey". In: *The Journal of Machine Learning Research* 20.1 (2019), pp. 1997–2017. DOI: 10.48550/arXiv.1808.05377.

Falk, Thorsten et al. "U-Net: deep learning for cell counting, detection, and morphometry". In: *Nature Methods* 16 (2019), pp. 67–70. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0261-2.

Faust, Oliver, Yuki Hagiwara, Tan Jen Hong, Oh Shu Lih, and U Rajendra Acharya. "Deep learning for healthcare applications based on physiological signals: A review". In: *Computer Methods and Programs in Biomedicine* 161 (2018), pp. 1–13. DOI: 10.1016/j.cmpb.2018.04.005.

Faust, Oliver, Hajar Razaghi, Ragab Barika, Edward J Ciaccio, and U Rajendra Acharya. "A review of automated sleep stage scoring based on physiological signals for the new millennia". In: *Computer Methods and Programs in Biomedicine* 176 (2019), pp. 81–91. DOI: 10.1016/j.cmpb.2019.04.032.

Feinberg, I and TC Floyd. "Systematic trends across the night in human sleep cycles". In: *Psychophysiology* 16.3 (1979), pp. 283–291. DOI: 10.1111/j.1469-8986.1979.tb02991.x.

Ferri, Raffaele, Silvia Miano, Oliviero Bruni, Jitka Vankova, Sona Nevsimalova, Stefano Vandi, Pasquale Montagna, Luigi Ferini-Strambi, and Giuseppe Plazzi. "NREM sleep alterations in narcolepsy/cataplexy". In: *Clinical neurophysiology* 116.11 (2005), pp. 2675–2684. DOI: 10.1016/j.clinph.2005.08.004.

Ferri, Raffaele, Marco Zucconi, Mauro Manconi, Oliviero Bruni, Silvia Miano, Giuseppe Plazzi, and Luigi Ferini-Strambi. "Computer-assisted detection of nocturnal leg motor activity in patients with restless legs syndrome and periodic leg movements during sleep". In: *Sleep* 28.8 (2005), pp. 998–1004. DOI: 10.1093/sleep/28.8.998.

Finnigan, Simon and Michel JAM van Putten. "EEG in ischaemic stroke: quantitative EEG can uniquely inform (sub-) acute prognoses and clinical management". In: *Clinical neurophysiology* 124.1 (2013), pp. 10–19. DOI: 10.1016/j.clinph.2012.07.003.

Fiorillo, Luigi, Giuliana Monachino, Julia van der Meer, Marco Pesce, Jan D Warncke, Markus H Schmidt, Claudio LA Bassetti, Athina Tzovara, Paolo Favaro, and Francesca D Faraci. "U-Sleep's resilience to AASM guidelines". In: *npj Digital Medicine* 6.1 (2023), p. 33. DOI: 10.1038/s41746-023-00784-0.

Fiorillo, Luigi, Alessandro Puiatti, Michela Papandrea, Pietro-Luca Ratti, Paolo Favaro, Corinne Roth, Panagiotis Bargiotas, Claudio L Bassetti, and Francesca D Faraci. "Automated sleep scoring: A review of the latest approaches". In: *Sleep medicine reviews* 48 (2019), p. 101204. DOI: 10.1016/j.smrv.2019.07.007.

French, Robert M. "Catastrophic forgetting in connectionist networks". In: *Trends in cognitive sciences* 3.4 (1999), pp. 128–135. DOI: 10.1002/0470018860.s00096.

Fukushima, Kunihiko. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". In: *Biological cybernetics* 36.4 (1980), pp. 193–202. DOI: 10.1007/BF00344251.

Gabor, Dennis. "Theory of communication. Part 1: The analysis of information". In: *Journal of the Institution of Electrical Engineers-part III: radio and communication engineering* 93.26 (1946), pp. 429–441. DOI: 10.1049/ji-3-2.1946.0074.

Gan, Hock S, Muhammad H Ramlee, Abdul W Wahab, Yee S Lee, and Atsushi Shimizu. "From classical to deep learning: Review on cartilage and bone segmentation techniques in knee osteoarthritis research". In: *Artif Intell Rev* 54 (2020), pp. 2445–2494. DOI: 10.1007/s10462-020-09924-4.

Ganaye, P., M. Sdika, and H. Benoit-Cattin. "Towards integrating spatial localization in convolutional neural networks for brain image segmentation". In: *International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2018, pp. 621–625. DOI: 10.1109/ISBI.2018.8363652.

Garbarino, Sergio, Paola Lanteri, Paolo Durando, Nicola Magnavita, and Walter G. Sannita. "Comorbidity, mortality, quality of life and the healthcare/welfare/social costs of disordered sleep: A rapid review". In: *International Journal of Environmental Research and Public Health* 13.8 (2016), p. 831. ISSN: 1660-4601. DOI: 10.3390/ijerph13080831.

Ghassemi, M. M., B. E. Moody, L. H. Lehman, C. Song, Q. Li, H. Sun, R. G. Mark, M. B. Westover, and G. D. Clifford. "You Snooze, You Win: the PhysioNet/Computing in Cardiology Challenge 2018". In: *2018 Computing in Cardiology Conference (CinC)*. Vol. 45. 2018, pp. 1–4. DOI: 10.22489/CinC.2018.049.

Giger, Maryellen L., Heang-Ping Chan, and John Boone. "Anniversary Paper: History and status of CAD and quantitative image analysis: The role of Medical Physics and AAPM". In: *Medical Physics* 35.12 (2008), pp. 5799–5820. DOI: https://doi.org/10.1118/1.3013555.

Goldberger, A. L., L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals". In: *Circulation* 101.23 (2000), e215–e220. DOI: 10.1161/01.CIR.101.23.e215.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Gorodkin, Jan. "Comparing two K-category assignments by a K-category correlation coefficient". In: *Computational biology and chemistry* 28.5-6 (2004), pp. 367–374. DOI: 10.1016/j.compbiolchem.2004.09.006.

Gramfort, Alexandre et al. "MEG and EEG Data Analysis with MNE-Python". In: *Frontiers in Neuroscience* 7.267 (2013), pp. 1–13. DOI: 10.3389/fnins.2013.00267.

Grigg-Damberger, Madeleine, David Gozal, Carole L Marcus, Stuart F Quan, Carol L Rosen, Ronald D Chervin, Merrill Wise, Daniel L Picchietti, Stephan H Sheldon, and Conrad Iber. "The visual scoring of sleep and arousal in infants and children". In: *Journal of Clinical Sleep Medicine* 3.02 (2007), pp. 201–240. DOI: 10.5664/jcsm.26819.

Grözinger, Michael, Joachim Röschke, and Bert Klöppel. "Automatic recognition of rapid eye movement (REM) sleep by artificial neural networks". In: *Journal of sleep research* 4.2 (1995), pp. 86–91. DOI: 10.1111/j.1365-2869.1995.tb00156.x.

Gu, Jiuxiang, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. "Recent advances in convolutional neural networks". In: *Pattern recognition* 77 (2018), pp. 354–377. DOI: 10.1016/j.patcog.2017.10.013.

Guillot, Antoine, Fabien Sauvet, Emmanuel H During, and Valentin Thorey. "Dreem Open Datasets: Multi-Scored Sleep Datasets to compare Human and Automated sleep staging". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28.9 (2019), pp. 1955–1965. DOI: 10.1109/TNSRE.2020.3011181.

Guo, Zhe, Xiang Li, Heng Huang, Ning Guo, and Quanzheng Li. "Deep Learning-Based Image Segmentation on Multimodal Medical Imaging". In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 3.2 (2019), pp. 162–169. DOI: 10.1109/TRPMS.2018.2890359.

Halász, P, M.G Terzano, and L Parrino. "Spike-wave discharge and the microstructure of sleep-wake continuum in idiopathic generalised epilepsy". In: *Neurophysiologie Clinique/Clinical Neurophysiology* 32.1 (2002), pp. 38–53. ISSN: 0987-7053. DOI: https://doi.org/10.1016/S0987-7053(01)00290-8.

Halász, Péter, Mario Terzano, Liborio Parrino, and Róbert Bódizs. "The nature of arousal in sleep". In: *Journal of sleep research* 13.1 (2004), pp. 1–23. DOI: 10.1111/j.1365-2869.2004.00388.x.

Hamidinekoo, Azam, Erika Denton, Andrik Rampun, Kate Honnor, and Reyer Zwiggelaar. "Deep learning in mammography and breast histology, an overview and future trends". In: *Medical image analysis* 47 (2018), pp. 45–67. DOI: 10.1016/j.media.2018.03.006.

Hasan, Joel. "Differentiation of normal and disturbed sleep by automatic analysis." In: *Acta physiologica scandinavica. supplementum* 526 (1983), pp. 1–103.

— "Past and future of computer-assisted sleep analysis and drowsiness assessment". In: *Journal of clinical neurophysiology* 13.4 (1996), pp. 295–313. DOI: 10.1097/00004691-199607000-00004.

Hasan, Joel, Kari Hirvonen, Alpo Värri, Veikko Häkkinen, and Pekka Loula. "Validation of computer analysed polygraphic patterns during drowsiness and sleep onset". In: *Electroencephalography and clinical neurophysiology* 87.3 (1993), pp. 117–127. DOI: 10.1016/0013-4694(93)90118-F.

Hatamizadeh, Ali, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images". In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I*. Springer. 2022, pp. 272–284. DOI: 10.1007/978-3-031-08999-2_22.

Hatamizadeh, Ali, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. "Unetr: Transformers for 3d medical image segmentation". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2022, pp. 574–584. DOI: 10.1109/WACV51458.2022.00181.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

— "Identity mappings in deep residual networks". In: *Computer Vision–ECCV 2016: 14th European Conference, Proceedings, Part IV 14*. Springer. 2016, pp. 630–645. DOI: 10.1007/978-3-319-46493-0_38.

He, Xin, Kaiyong Zhao, and Xiaowen Chu. "AutoML: A survey of the state-of-the-art". In: *Knowledge-Based Systems* 212 (2021), p. 106622. DOI: 10.1016/j.knosys.2020.106622.

He, Yufan, Dong Yang, Holger Roth, Can Zhao, and Daguang Xu. "Dints: Differentiable neural network topology search for 3d medical image segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 5841–5850. DOI: 10.1109/CVPR46437.2021.00578.

Hesamian, Mohammad Hesam, Wenjing Jia, Xiangjian He, and Paul Kennedy. "Deep learning techniques for medical image segmentation: achievements and challenges". In: *Journal of digital imaging* 32 (2019), pp. 582–596. DOI: 10.1007/s10278-019-00227-x.

Hinton, Geoffrey E, Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets". In: *Neural computation* 18.7 (2006), pp. 1527–1554. DOI: 10.1162/neco.2006.18.7.1527.

Hochreiter, Sepp and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Hofmann, Heike, Hadley Wickham, and Karen Kafadar. "Letter-Value Plots: Boxplots for Large Data". In: *Journal of Computational and Graphical Statistics* 26.3 (2017), pp. 469–477. DOI: 10.1080/10618600.2017.1305277.

Hubel, David H and Torsten N Wiesel. "Receptive fields and functional architecture of monkey striate cortex". In: *The Journal of physiology* 195.1 (1968), pp. 215–243. DOI: 10.1113/jphysiol.1968.sp008455.

Hutter, Frank, Lars Kotthoff, and Joaquin Vanschoren. *Automated machine learning: methods, systems, challenges*. Springer Nature, 2019. DOI: 10.1007/978-3-030-05318-5.

Iber, Conrad and AASM. *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*. Westchester, IL: American Academy of Sleep Medicine, 2007.

ICRU. "ICRU Report 62, Prescribing, Recording and Reporting Photon Beam Therapy (Supplement to ICRU Report 50)". In: *ICRU News* (1999).

Ioffe, Sergey and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *International Conference on Machine Learning (ICML)*. PMLR, 2015, pp. 448–456. DOI: 10.48550/arXiv.1502.03167.

Isensee, Fabian et al. "nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation". In: *CoRR* abs/1809.10486 (2018). DOI: 10.1007/978-3-658-25326-4_7.

Jasper, Herbert H. "Report of the committee on methods of clinical examination in electroencephalography: 1957". In: *Electroencephalogr Clin Neurophysiol* 10 (1958), pp. 370–375.

Jennum, P. and R. L. Riha. "Epidemiology of sleep apnoea/hypopnoea syndrome and sleep-disordered breathing". In: *European Respiratory Journal* 33.4 (2009), pp. 907–914. ISSN: 0903-1936. DOI: 10.1183/09031936.00180108.

Jiang, Zeyu, Changxing Ding, Minfeng Liu, and Dacheng Tao. "Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task". In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I 5*. Springer. 2020, pp. 231–241. DOI: 10.1007/978-3-030-46640-4_22.

Jordan, Kenneth G. "Emergency EEG and continuous EEG monitoring in acute ischemic stroke". In: *Journal of clinical neurophysiology* 21.5 (2004), pp. 341–352. DOI: 10.1097/01.WNP.0000145005.59766.D2.

Joskowicz, Leo, D. Cohen, N. Caplan, and J. Sosna. "Inter-observer variability of manual contour delineation of structures in CT". In: *European Radiology* 29.3 (2019), pp. 1391–1399. ISSN: 1432-1084. DOI: 10.1007/s00330-018-5695-5.

Ju, Yo-El, Brendan Lucey, and David M Holtzman. "Sleep and Alzheimer disease pathology — a bidirectional relationship". In: *Nature Reviews Neurology* 10.2 (2014), pp. 115–119. DOI: 10.1038/nrneurol.2013.269.

Kales, Anthony. and Allan. Rechtschaffen. *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. U. S. National Institute of Neurological Diseases and Blindness, Neurological Information Network Bethesda, Md, 1968, p. 57.

Kashyap, Snehil, Hong Zhang, Karthik R Rao, and Milan Sonka. "Learning-based cost functions for 3-D and 4-D multi-surface multi-object segmentation of knee MRI: Data from the osteoarthritis initiative". In: *IEEE Trans Med Imaging* 37 (2018), pp. 1103–1113. DOI: 10.1109/TMI.2017.2781541.

Kaushal, Amit, Russ Altman, and Curt Langlotz. "Geographic distribution of US cohorts used to train deep learning algorithms". In: *Jama* 324.12 (2020), pp. 1212–1213. DOI: 10.1001/jama.2020.12067.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. "LightGBM: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems* 30 (2017).

Kellgren, John and Jeffrey Lawrence. "Radiological assessment of osteo-arthrosis". In: *Ann Rheum Dis* 16 (1957), pp. 494–502. DOI: 10.1136/ard.16.4.494.

Kemp, B., A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberye. "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG". In: *IEEE Transactions on Biomedical Engineering* 47.9 (2000), pp. 1185–1194. DOI: 10.1109/10.867928.

Kemp, Bob. "A proposal for computer-based sleep/wake analysis". In: *Journal of sleep research* 2.3 (1993), pp. 179–185. DOI: 10.1111/j.1365-2869.1993.tb00084.x.

Kerkhof, Gerard A. "Epidemiology of sleep and sleep disorders in The Netherlands". In: *Sleep medicine* 30 (2017), pp. 229–239. DOI: 10.1016/j.sleep.2016.09.015.

Khalighi, Sirvan, Teresa Sousa, José Moutinho dos Santos, and Urbano Nunes. "ISRUC-Sleep: A comprehensive public dataset for sleep researchers." In: *Computer Methods and Programs in Biomedicine* 124 (2016), pp. 180–192. DOI: 10.1016/j.cmpb.2015.10.013.

Kheirandish-Gozal, Leila, Silvia Miano, Oliviero Bruni, Raffaele Ferri, Jacopo Pagani, Maria Pia Villa, and David Gozal. "Reduced NREM sleep instability in children with sleep disordered breathing". In: *Sleep* 30.4 (2007), pp. 450–457. DOI: 10.1093/sleep/30.4.450.

Kim, Young do, Masayoshi Kurachi, Motoshi Horita, Kohki Matsuura, and Yasuko Kamikawa. "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". In: *Psychiatry and Clinical Neurosciences* 47.1 (1993), pp. 91–97. DOI: https://doi.org/10.1111/j.1440-1819.1993.tb02035.x.

Kingma, Diederik P. and Jimmy Lei Ba. "Adam: A method for stochastic optimization". In: *International Conference on Learning Representations (ICLR)*. 2015. DOI: 10.48550/arXiv.1412.6980.

Klosh, G. et al. "The SIESTA project polygraphic and clinical database". In: *IEEE Engineering in Medicine and Biology Magazine* 20.3 (2001), pp. 51–57. DOI: 10.1109/51.932725.

Koch, Henriette, Poul Jennum, and Julie A. E. Christensen. "Automatic sleep classification using adaptive segmentation reveals an increased number of rapid eye movement sleep transitions". In: *Journal of Sleep Research* 28.2 (2019), e12780. DOI: 10.1111/jsr.12780.

Koch, Henriette, Logan Douglas Schneider, Laurel A. Finn, Eileen B. Leary, Paul E. Peppard, Erika Hagen, Helge Bjarup Dissing Sorensen, Poul Jennum, and Emmanuel Mignot. "Breathing Disturbances Without Hypoxia Are Associated With Objective Sleepiness in Sleep Apnea". In: *Sleep* 40.11 (2017), zsx152. ISSN: 1550-9109. DOI: 10.1093/sleep/zsx152.

Koch, Thorbjørn, Mathias Perslev, Christian Igel, and Sami Brand. "Accurate Segmentation of Dental Panoramic Radiographs with U-Nets". In: *IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE Press, 2019, pp. 15–19. DOI: 10.1109/ISBI.2019.8759563.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey Hinton. "Imagenet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (2017), pp. 84–90. DOI: 10.1145/3065386.

Krueger, James, Joseph Nguyen, Cheryl Dykstra-Aiello, and Ping Taishi. "Local sleep". In: *Sleep medicine reviews* 43 (2019), pp. 14–21. DOI: 10.1016/j.smrv.2018.10.001.

Krueger, James M, David M Rector, Sandip Roy, Hans PA Van Dongen, Gregory Belenky, and Jaak Panksepp. "Sleep as a fundamental property of neuronal assemblies". In: *Nature Reviews Neuroscience* 9.12 (2008), pp. 910–919. DOI: 10.1038/nrn2521.

Kubicki, St and WM Herrmann. "The future of computer-assisted investigation of the polysomnogram: sleep microstructure". In: *Journal of Clinical Neurophysiology* 13.4 (1996), pp. 285–294. DOI: 10.1097/00004691-199607000-00003.

Kubicki, St., L. Höller, I. Berg, C. Pastelak-Price, and R. Dorow. "Sleep EEG Evaluation: A Comparison of Results Obtained by Visual Scoring and Automatic Analysis with the Oxford Sleep Stager". In: *Sleep* 12.2 (1989), pp. 140–149. ISSN: 0161-8105. DOI: 10.1093/sleep/12.2.140.

Kullback, Solomon and Richard A Leibler. "On information and sufficiency". In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86. DOI: 10.1214/aoms/1177729694.

Kuo, Chih En and Guan Ting Chen. "Automatic Sleep Staging Based on a Hybrid Stacked LSTM Neural Network: Verification Using Large-Scale Dataset". In: *IEEE Access* 8 (2020), pp. 111837–111849. DOI: 10.1109/ACCESS.2020.3002548.

Lakhani, Paras and Baskaran Sundaram. "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks". In: *Radiology* 284.2 (2017), pp. 574–582. DOI: 10.1148/radiol.2017162326.

Landman, B.A., A. Ribbens, B. Lucas, C.D. Christos, B. Avants, C. Ledig, D. Ma, D. Rueckert, S.K. Warfield, D. Vandermeulen, et al. *MICCAI 2012 Workshop on Multi-Atlas Labeling*. CreateSpace Independent Publishing Platform, 2012. ISBN: 978-1-4791-2618-7.

Lauritzen, Andreas D., Alejandro Rodríguez-Ruiz, My Catarina von Euler-Chelpin, Elsebeth Lynge, Ilse Vejborg, Mads Nielsen, Nico Karssemeijer, and Martin Lillholm. "An Artificial Intelligence–based Mammography Screening Protocol for Breast Cancer: Outcome and Radiologist Workload". In: *Radiology* 304.1 (2022). PMID: 35438561, pp. 41–49. DOI: 10.1148/radiol.210948.

Lea, Colin, René Vidal, Austin Reiter, and Gregory D. Hager. "Temporal Convolutional Networks: A Unified Approach to Action Segmentation". In: *CoRR* abs/1608.08242 (2016). DOI: 10.1007/978-3-319-49409-8_7.

Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.

LeCun, Yann, Y Bengio, and Geoffrey Hinton. "Deep Learning". In: *Nature* 521.7553 (May 2015), pp. 436–444. DOI: 10.1038/nature14539.

LeCun, Yann, Leon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. "Efficient BackProp". In: *Neural Networks: Tricks of the Trade*. Ed. by Genevieve B. Orr and Klaus-Robert Müller. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 9–50. ISBN: 978-3-540-49430-0. DOI: 10.1007/3-540-49430-8_2.

Lee, Harlin, Boyue Li, Shelly DeForte, Mark L. Splaingard, Yungui Huang, Yuejie Chi, and Simon L. Linwood. "A large collection of real-world pediatric sleep studies". In: *Scientific Data* 9.1 (2022), p. 421. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01545-6.

Lee, Hyungjin, Hyojin Hong, and Jinwook Kim. "BCD-NET: A novel method for cartilage segmentation of knee MRI via deep segmentation networks with bone-cartilage-complex modeling". In:

Proceedings - IEEE 15th International Symposium on Biomedical Imaging (ISBI). The Institute of Electrical and Electronics Engineers. 2018, pp. 1538–1541. DOI: 10.1109/ISBI.2018.8363866.

Li, Jun, Junyu Chen, Yucheng Tang, Ce Wang, Bennett A. Landman, and S. Kevin Zhou. "Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives". In: *Medical Image Analysis* 85 (2023), p. 102762. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2023.102762.

Li, Wenqi, Guotai Wang, Lucas Fidon, Sebastien Ourselin, M Jorge Cardoso, and Tom Vercauteren. "On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task". In: *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings 25*. Springer. 2017, pp. 348–360. DOI: 10.1007/978-3-319-59050-9_28.

Litjens, Geert, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. "A survey on deep learning in medical image analysis". In: *Medical Image Analysis* 42 (2017), pp. 60–88. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2017.07.005.

Liu, Fang, Zhaoye Zhou, Hyungseok Jang, Alexey Samsonov, Guang Zhao, and Richard Kijowski. "Deep convolutional neural network and 3D deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging". In: *Magn Reson Med* 79 (2018), pp. 2379–2391. DOI: 10.1002/mrm.26841.

Liu, Shengfeng, Yi Wang, Xin Yang, Baiying Lei, Li Liu, Shawn Xiang Li, Dong Ni, and Tianfu Wang. "Deep learning in medical ultrasound analysis: a review". In: *Engineering* 5.2 (2019), pp. 261–275. DOI: 10.1109/ACCESS.2021.3071301.

Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022. DOI: 10.1109/ICCV48922.2021.00986.

Liu, Zhuang, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. "A convnet for the 2020s". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 11976–11986. DOI: 10.1109/CVPR52688.2022.01167.

Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully Convolutional Networks for Semantic Segmentation". In: *Computer Vision and Pattern Recognition (CVPR)*. Vol. abs/1411.4038. IEEE, 2014, pp. 3431–3440. DOI: 10.1109/TPAMI.2016.2572683.

Luck, Steven J. *An introduction to the event-related potential technique*. MIT press, 2014.

Lundervold, Alexander Selvikvåg and Arvid Lundervold. "An overview of deep learning in medical imaging focusing on MRI". In: *Zeitschrift für Medizinische Physik* 29.2 (2019), pp. 102–127. DOI: 10.1016/j.zemedi.2018.11.002.

Luo, Wenjie, Yujia Li, Raquel Urtasun, and Richard Zemel. *Understanding the Effective Receptive Field in Deep Convolutional Neural Networks*. 2017. DOI: 10.48550/ARXIV.1701.04128.

Maintz, JB Antoine and Max A Viergever. "A survey of medical image registration". In: *Medical image analysis* 2.1 (1998), pp. 1–36. DOI: 10.1016/S1361-8415(01)80026-8.

Marcus, Carole L. et al. "A randomized trial of adenotonsillectomy for childhood sleep apnea". In: *New England Journal of Medicine* 368 (2013), pp. 2366–2376. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1215881.

Marcus, Daniel S, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. "Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults". In: *Journal of Cognitive Neuroscience* 19.9 (2007), pp. 1498–1507. DOI: 10.1162/jocn.2007.19.9.1498.

Martin, WB, LC Johnson, SS Viglione, P Naitoh, RD Joseph, and JD Moses. "Pattern recognition of EEG-EOG as a technique for all-night sleep stage scoring". In: *Electroencephalography and clinical neurophysiology* 32.4 (1972), pp. 417–427. DOI: 10.1016/0013-4694(72)90009-0.

Mascetti, Gian Gastone. "Unihemispheric sleep and asymmetrical sleep: behavioral, neurophysiological, and functional perspectives". In: *Nature and Science of Sleep* (2016), pp. 221–238.

Masegosa, Andrés R., Stephan S Lorenzen, Christian Igel, and Yevgeny Seldin. "Second Order PAC-Bayesian Bounds for the Weighted Majority Vote". In: *Advances in Neural Information Processing Systems (NeurIPS) 33*. 2020.

Mathers, Colin D and Dejan Loncar. "Projections of global mortality and burden of disease from 2002 to 2030". In: *PLoS medicine* 3.11 (2006), e442. DOI: 10.1371/journal.pmed.0030442.

Menze, Bjoern H, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. "The multimodal brain tumor image segmentation benchmark (BRATS)". In: *IEEE transactions on medical imaging* 34.10 (2014), pp. 1993–2024. DOI: 10.1109/TMI.2014.2377694.

Miller, Christopher B. et al. "Clusters of insomnia disorder: An exploratory cluster analysis of objective sleep parameters reveals differences in neurocognitive functioning, quantitative EEG, and heart rate variability". In: *Sleep* 39.11 (2016), pp. 1993–2004. ISSN: 1550-9109. DOI: 10.5665/sleep.6230.

Milletari, Fausto, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation". In: *2016 fourth international conference on 3D vision (3DV)*. Ieee. 2016, pp. 565–571. DOI: 10.1109/3DV.2016.79.

Moeskops, Pim, Max A. Viergever, Adriënne M. Mendrik, Linda S. de Vries, Manon J. N. L. Benders, and Ivana Isgum. "Automatic Segmentation of MR Brain Images With a Convolutional Neural Network". In: *IEEE Transactions on Medical Imaging* 35.5 (2016), pp. 1252–1261. DOI: 10.1109/TMI.2016.2548501.

Moghbel, Mehrdad, Syamsiah Mashohor, Rozi Mahmud, and M Iqbal Bin Saripan. "Review of liver segmentation and computer assisted detection/diagnosis methods in computed tomography". In: *Artificial Intelligence Review* 50 (2018), pp. 497–537. DOI: 10.1007/s10462-017-9550-x.

Mousavi, Sajad, Fatemeh Afghah, and U. Rajendra Acharya. "SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach". In: *PLOS ONE* 14.5 (2019), pp. 1–15. DOI: 10.1371/journal.pone.0216456.

Mukhametov, LM, Ya Supin, and IG Polyakova. "Interhemispheric asymmetry of the electroencephalographic sleep patterns in dolphins." In: *Brain research* (1977). DOI: 10.1016/0006-8993(77)90835-6.

Mulert, Christoph. "Simultaneous EEG and fMRI: towards the characterization of structure and dynamics of brain networks". In: *Dialogues in clinical neuroscience* (2022). DOI: 10.31887/DCNS.2013.15.3/cmulert.

Myerson, Saul G, Jane Francis, and Stefan Neubauer. *Cardiovascular magnetic resonance*. Oxford University Press, 2010. DOI: 10.1093/med/9780199549573.001.1.

Nevitt, M, D Felson, and Gayle Lester. "The osteoarthritis initiative". In: *Protocol for the cohort study* 1 (2006).

Nikkonen, Sami, Henri Korkalainen, Akseli Leino, Sami Myllymaa, Brett Duce, Timo Leppänen, and Juha Töyräs. "Automatic respiratory event scoring in obstructive sleep apnea using a long short-term memory neural network". In: *IEEE Journal of Biomedical and Health Informatics* 25.8 (2021), pp. 2917–2927. DOI: 10.1109/JBHI.2021.3064694.

Nir, Yuval, RichardáJ Staba, Thomas Andrillon, VladyslaváV Vyazovskiy, Chiara Cirelli, Itzhak Fried, and Giulio Tononi. "Regional slow waves and spindles in human sleep". In: *Neuron* 70.1 (2011), pp. 153–169. DOI: 10.1016/j.neuron.2011.02.043.

Norman, Benjamin, Valentina Pedoia, and Sharmila Majumdar. "Use of 2D U-net convolutional neural networks for automated cartilage and meniscus segmentation of knee MR imaging data to determine relaxometry and morphometry". In: *Radiology* 288 (2018), pp. 177–185. DOI: 10.1148/radiol.2018172322.

O'Reilly, Christian, Nadia Gosselin, Julie Carrier, and Tore Nielsen. "Montreal Archive of Sleep Studies: an open-access resource for instrument benchmarking and exploratory research". In: *Journal of Sleep Research* 23.6 (2014), pp. 628–635. ISSN: 1365-2869. DOI: 10.1111/jsr.12169.

Odena, Augustus, Vincent Dumoulin, and Chris Olah. "Deconvolution and Checkerboard Artifacts". In: *Distill* (2016). DOI: 10.23915/distill.00003.

Olesen, Alexander Neergaard, Matteo Cesari, Julie Anja Engelhard Christensen, Helge Bjarup Dissing Sorensen, Emmanuel Mignot, and Poul Jennum. "A comparative study of methods for automatic detection of rapid eye movement abnormal muscular activity in narcolepsy". In: *Sleep Medicine* 44 (2018), pp. 97–105. ISSN: 1878-5506. DOI: 10.1016/j.sleep.2017.11.1141.

Oliveira, Francisco PM and Joao Manuel RS Tavares. "Medical image registration: a review". In: *Computer methods in biomechanics and biomedical engineering* 17.2 (2014), pp. 73–93. DOI: 10.1080/10255842.2012.670855.

Opbroek, Annegreet van, M. Arfan Ikram, Meike W. Vernooij, and Marleen de Bruijne. "Transfer Learning Improves Supervised Image Segmentation Across Imaging Protocols". In: *IEEE Transactions on Medical Imaging* 34.5 (2015), pp. 1018–1030. DOI: 10.1109/TMI.2014.2366792.

Pal, Nikhil R and Sankar K Pal. "A review on image segmentation techniques". In: *Pattern Recognition* 26.9 (1993), pp. 1277–1294. ISSN: 0031-3203. DOI: https://doi.org/10.1016/0031-3203(93)90135-J.

Panfilov, Egor, Aleksei Tiulpin, Stefan Klein, Miika Nieminen, and Simo Saarakkala. "Improving robustness of deep learning based knee mri segmentation: Mixup and adversarial domain adaptation". In: *Proceedings - IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. The Institute of Electrical and Electronics Engineers. 2019, pp. 450–459. DOI: 10.1109/ICCVW.2019.00057.

Parikh, Ravi B, Ziad Obermeyer, and Amol S Navathe. "Regulation of predictive analytics in medicine". In: *Science* 363.6429 (2019), pp. 810–812. DOI: 10.1126/science.aaw0029.

Parrino, Liborio, Mirella Boselli, Maria Cristina Spaggiari, Arianna Smerieri, and Mario Giovanni Terzano. "Cyclic alternating pattern (CAP) in normal sleep: polysomnographic parameters in different age groups". In: *Electroencephalography and Clinical Neurophysiology* 107.6 (1998), pp. 439–450. ISSN: 0013-4694. DOI: https://doi.org/10.1016/S0013-4694(98)00108-4.

Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks". In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta and David McAllester. Vol. 28. Proceedings of Machine Learning Research 3. Atlanta, Georgia, USA: PMLR, 2013, pp. 1310–1318. DOI: 10.48550/arXiv.1211.5063.

Patel, Aakash K, Vamsi Reddy, and John F Araujo. "Physiology, sleep stages". In: *StatPearls [Internet]*. StatPearls Publishing, 2022.

Pedregosa, F. et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830. DOI: 10.48550/arXiv.1201.0490.

Peng, Peng, Karim Lekadir, Ali Gooya, Ling Shao, Steffen E Petersen, and Alejandro F Frangi. "A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging". In: *Magnetic Resonance Materials in Physics, Biology and Medicine* 29 (2016), pp. 155–195. DOI: 10.1007/s10334-015-0521-4.

Penzel, T, K Stephan, S Kubicki, and WM Herrmann. "Integrated sleep analysis, with emphasis on automatic methods". In: *Epilepsy research. Supplement* 2 (1991), pp. 177–204.

Penzel, Thomas and Regina Conradt. "Computer based sleep recording and analysis". In: *Sleep medicine reviews* 4.2 (2000), pp. 131–148. DOI: 10.1053/smrv.1999.0087.

Perslev, Mathias, Erik B. Dam, Akshay Pai, and Christian Igel. "One Network To Segment Them All: A General, Lightweight System for Accurate 3D Medical Image Segmentation". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. LNCS 11765. Springer, 2019, pp. 30–38. DOI: 10.1007/978-3-030-32245-8_4.

Perslev, Mathias, Sune Darkner, Lykke Kempfner, Miki Nikolic, Poul Jennum, and Christian Igel. "U-Sleep: resilient high-frequency sleep staging". In: *npj Digital Medicine* 4.1 (Apr. 15, 2021). ISSN: 2398-6352. DOI: 10.1038/s41746-021-00440-5.

Perslev, Mathias, Michael Jensen, Sune Darkner, Poul Jørgen Jennum, and Christian Igel. "U-Time: A Fully Convolutional Network for Time Series Segmentation Applied to Sleep Staging". In: *Advances in Neural Information Processing Systems (NeurIPS) 32*. 2019, pp. 4415–4426. DOI: 10.48550/arXiv.1910.11162.

Perslev, Mathias, Akshay Pai, Jos Runhaar, Christian Igel, and Erik B. Dam. "Cross-Cohort Automatic Knee MRI Segmentation With Multi-Planar U-Nets". In: *Journal of Magnetic Resonance Imaging* 55.6 (2022), pp. 1650–1663. DOI: 10.1002/jmri.27978.

Perslev, Mathias, Anders Sode West, Sofie Amalie Simonsen, Laura Bødker Ponsaing, Helle Klingenberg Iversen, Christian Igel, and Poul Jørgen Jennum. "Automatic detection of abnormal sleeping patterns in stroke patients using high-frequency sleep staging". In: *Journal of Sleep Research* 31.S1 (2022), O130/P964. DOI: https://doi.org/10.1111/jsr.13739.

Pfurtscheller, G, D Flotzinger, and K Matuschik. "Sleep Classification in Infants Based on Artificial Neural Networks. Schlafklassifikation mit Hilfe neuronaler Netzwerke". In: (1992). DOI: 10.1515/bmte.1992.37.6.122.

Pham, Dzung L, Chenyang Xu, and Jerry L Prince. "Current methods in medical image segmentation". In: *Annual review of biomedical engineering* 2.1 (2000), pp. 315–337. DOI: 10.1146/annurev.bioeng.2.1.315.

Phan, Huy, Fernando Andreotti, Navin Cooray, Oliver Y. Chen, and Maarten De Vos. "SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep

Staging". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27.3 (2019), pp. 400–410. ISSN: 1534-4320. DOI: 10.1109/TNSRE.2019.2896659.

Phan, Huy, Fernando Andreotti, Navin Cooray, Oliver Y. Chén, and Maarten De Vos. "Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification". In: *CoRR* abs/1805.06546 (2018). DOI: 10.1109/TBME.2018.2872652.

Phan, Huy, Oliver Y. Chén, Philipp Koch, Zongqing Lu, I. Mcloughlin, A. Mertins, and M. Vos. "Towards More Accurate Automatic Sleep Staging via Deep Transfer Learning". In: *IEEE Transactions on Biomedical Engineering* (2020). DOI: 10.1109/TBME.2020.3020381.

Phan, Huy, Oliver Y. Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. *XSleepNet: Multi-View Sequential Model for Automatic Sleep Staging*. 2020. DOI: 10.1109/TPAMI.2021.3070057.

Phan, Huy, Kaare Mikkelsen, Oliver Y Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. "Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification". In: *IEEE Transactions on Biomedical Engineering* 69.8 (2022), pp. 2456–2467. DOI: 10.1109/TBME.2022.3147187.

Ponsaing, Laura B., Helle K. Iversen, and Poul Jennum. "Polysomnographic indicators of mortality in stroke patients". In: *Sleep and Breathing* 21.2 (2017), pp. 235–242. ISSN: 1522-1709. DOI: 10.1007/s11325-016-1387-z.

Prasoon, Adhish, Kersten Petersen, Christian Igel, François Lauze, Erik B. Dam, and Mads Nielsen. "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16*. Vol. 8150. LNCS. Springer. Springer, 2013, pp. 246–253. DOI: 10.1007/978-3-642-40763-5_31.

Priano, Lorenzo, Matteo Bigoni, Giovanni Albani, Luigi Sellitti, Emanuela Giacomotti, Roberto Picconi, Riccardo Cremascoli, Maurizio Zibetti, Leonardo Lopiano, and Alessandro Mauro. "Sleep microstructure in Parkinson's disease: cycling alternating pattern (CAP) as a sensitive marker of early NREM sleep instability". In: *Sleep Medicine* 61 (2019), pp. 57–62. DOI: 10.1016/j.sleep.2019.03.025.

Principe, Jose C and AMP Tome. "Performance and training strategies in feedforward neural networks: an application to sleep scoring". In: *Int Joint Conf Neural Networks, Washington DC*. Vol. 1. 1989, pp. 341–346. DOI: 10.1109/IJCNN.1989.118606.

Quan, Stuart et al. "The Sleep Heart Health Study: Design, Rationale, and Methods". In: *Sleep* 20.12 (1998), pp. 1077–85. ISSN: 0161-8105. DOI: 10.1093/sleep/20.12.1077.

Raj, Aadarsh, Sindhu Vishwanathan, Bhavesh Ajani, Karthik Krishnan, and Harsh Agarwal. "Automatic knee cartilage segmentation using fully volumetric convolutional neural networks for evaluation of osteoarthritis". In: *Proceedings - IEEE 15th International Symposium on Biomedi-*

cal Imaging (ISBI). The Institute of Electrical and Electronics Engineers. 2018, pp. 851–854. DOI: 10.1109/ISBI.2018.8363705.

Razzak, Muhammad Imran, Saeeda Naz, and Ahmad Zaib. "Deep learning for medical image processing: Overview, challenges and the future". In: *Classification in BioApps: Automation of Decision Making* (2018), pp. 323–350. DOI: 10.1007/978-3-319-65981-7_12.

Reader, Andrew J, Guillaume Corda, Abolfazl Mehranian, Casper da Costa-Luis, Sam Ellis, and Julia A Schnabel. "Deep learning for PET image reconstruction". In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 5.1 (2020), pp. 1–25. DOI: 10.1109/TRPMS.2020.3014786.

Real, Esteban, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Jie Tan, Quoc V Le, and Alexey Kurakin. "Large-scale evolution of image classifiers". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 2902–2911. DOI: 10.48550/arXiv.1703.01041.

Redline, Susan, Raouf Amin, et al. "The Childhood Adenotonsillectomy Trial (CHAT): Rationale, Design, and Challenges of a Randomized Controlled Trial Evaluating a Standard Surgical Procedure in a Pediatric Population". In: *Sleep* 34 (Nov. 2011), pp. 1509–17. DOI: 10.5665/sleep.1388.

Redline, Susan, Peter V. Tishler, Tor D. Tosteson, John Williamson, Kenneth Kump, Ilene Browner, Veronica Ferrette, and Patrick Krejci. "The familial aggregation of obstructive sleep apnea". In: *American Journal of Respiratory and Critical Care Medicine* 151 (3 1995), pp. 682–687. ISSN: 1073-449X. DOI: 10.1164/ajrccm.151.3.7881656.

Remeseiro, Beatriz and Veronica Bolon-Canedo. "A review of feature selection methods in medical applications". In: *Computers in Biology and Medicine* 112 (2019), p. 103375. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2019.103375.

Robert, Claude, Christian Guilpin, and Aymé Limoge. "Review of neural network applications in sleep research". In: *Journal of Neuroscience Methods* 79.2 (1998), pp. 187–193. DOI: 10.1016/S0165-0270(97)00178-7.

Roessler, R, F Collins, and R Ostman. "A period analysis classification of sleep stages". In: *Electroencephalography and clinical neurophysiology* 29.4 (1970), pp. 358–362. DOI: 10.1016/0013-4694(70)90043-X.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Vol. 9351. LNCS. Springer, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

Ronzhina, Marina, Oto Janoušek, Jana Kolářová, Marie Nováková, Petr Honzík, and Ivo Provazník. "Sleep scoring using artificial neural networks". In: *Sleep Medicine Reviews* 16.3 (2012), pp. 251–263. DOI: 10.1016/j.smrv.2011.06.003.

Rosen, Carol L., Dennis Auckley, Ruth Benca, Nancy Foldvary-Schaefer, Conrad Iber, Vishesh Kapur, Michael Rueschman, Phyllis Zee, and Susan Redline. "A multisite randomized trial of portable sleep studies and positive airway pressure autotitration versus laboratory-based polysomnography for the diagnosis and treatment of obstructive sleep apnea: The HomePAP study". In: *Sleep* 35.6 (2012), pp. 757–677. ISSN: 0161-8105. DOI: 10.5665/sleep.1870.

Rosen, Carol L., Emma K. Larkin, H. Lester Kirchner, Judith L. Emancipator, Sarah F. Bivins, Susan A. Surovec, Richard J. Martin, and Susan Redline. "Prevalence and risk factors for sleep-disordered breathing in 8- to 11-year-old children: Association with race and prematurity". In: *Journal of Pediatrics* 142.4 (2003), pp. 383–389. ISSN: 0022-3476. DOI: 10.1067/mpd.2003.28.

Rosenberg, Richard S. and Steven Van Hout. "The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring". In: *Journal of Clinical Sleep Medicine* 09.01 (2013), pp. 81–87. ISSN: 1550-9389, 1550-9397. DOI: 10.5664/jcsm.2350.

Roth, Holger, Chen Shen, Hirohisa Oda, Takaaki Sugino, Masahiro Oda, Yuichiro Hayashi, Kazunari Misawa, and Kensaku Mori. "A multi-scale pyramid of 3D fully convolutional networks for abdominal multi-organ segmentation". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*. Springer. 2018, pp. 417–425. DOI: 10.1007/978-3-030-00937-3_48.

Roth, Holger R., Le Lu, Jiamin Liu, Jianhua Yao, Ari Seff, Kevin Cherry, Lauren Kim, and Ronald M. Summers. "Improving Computer-Aided Detection Using Convolutional Neural Networks and Random View Aggregation". In: *IEEE Transactions on Medical Imaging* 35.5 (2016), pp. 1170–1181. DOI: 10.1109/TMI.2015.2482920.

Roy, Abhijit Guha, Sailesh Conjeti, Nassir Navab, and Christian Wachinger. "QuickNAT: Segmenting MRI Neuroanatomy in 20 seconds". In: *CoRR* abs/1801.04161 (2018).

Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature machine intelligence* 1.5 (2019), pp. 206–215. DOI: 10.1038/s42256-019-0048-x.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536. DOI: 10.1038/323533a0.

Runhaar, Jos, Marienke Van Middelkoop, Max Reijman, and et al. "Prevention of knee osteoarthritis in overweight females: The first preventive randomized controlled trial in osteoarthritis". In: *Am J Med* 128 (2015), 888–895.e4. DOI: 10.1016/j.joca.2012.02.551.

Sateia, Michael J. "International classification of sleep disorders-third edition: highlights and modifications". In: *Chest* 146.5 (2014), pp. 1387–1394. DOI: 10.1378/chest.14-0970.

Schenck, C.H. et al. "Corrigendum to "Rapid eye movement sleep behavior disorder: devising controlled active treatment studies for symptomatic and neuroprotective therapy—a consensus statement from the International Rapid Eye Movement Sleep Behavior Disorder Study Group" [Sleep Med 14(8) (2013) 795–806]". In: *Sleep Medicine* 15.1 (2014), p. 157. DOI: 10.1016/j.sleep.2013.11.001.

Schmidhuber, Jürgen. "Deep Learning in Neural Networks: An Overview". In: *CoRR* abs/1404.7828 (2015), pp. 85–117. DOI: 10.1016/j.neunet.2014.09.003.

Seim, Heike, Dagmar Kainmüller, Hans Lamecker, Markus Bindernagel, Jörg Malinowski, and Stefan Zachow. "Model-based auto-segmentation of knee bones and cartilage in MRI data". In: *Proceedings - Medical Image Analysis for the Clinic: a Grand Challenge in Conjunction with MICCAI*. Vol. 6361. Lecture Notes in Computer Science. Springer, Cham, 2010, pp. 215–223.

Shen, Dinggang, Guorong Wu, Suk Il, and Heung. "Deep learning in medical image analysis". In: *Annu Rev Biomed Eng* 19 (2017), pp. 221–248. DOI: 10.1146/annurev-bioeng-071516-044442.

Shiina, Shuichiro, Koki Sato, Ryosuke Tateishi, Motonori Shimizu, Hideko Ohama, Takeshi Hatanaka, Masashi Takawa, Hiroaki Nagamatsu, Yasuharu Imai, et al. "Percutaneous ablation for hepatocellular carcinoma: comparison of various ablation techniques and surgery". In: *Canadian Journal of Gastroenterology and Hepatology* 2018 (2018). DOI: 10.1155/2018/4756147.

Shorten, Connor and Taghi M Khoshgoftaar. "A survey on image data augmentation for deep learning". In: *Journal of big data* 6.1 (2019), pp. 1–48. DOI: 10.1186/s40537-019-0197-0.

Siclari, Francesca and Giulio Tononi. "Local aspects of sleep and wakefulness". In: *Current Opinion in Neurobiology* 44 (2017), pp. 222–227. DOI: 10.1016/j.conb.2017.05.008.

Silber, Michael H, Sonia Ancoli-Israel, Michael H Bonnet, Sudhansu Chokroverty, Madeleine M Grigg-Damberger, Max Hirshkowitz, Sheldon Kapen, Sharon A Keenan, Meir H Kryger, Thomas Penzel, et al. "The visual scoring of sleep in adults". In: *Journal of clinical sleep medicine* 3.02 (2007), pp. 121–131. DOI: 10.5664/jcsm.26814.

Silva, Andressa Alves da, Renato Gorga Bandeira de Mello, Camila Wohlgemuth Schaan, Flávio D Fuchs, Susan Redline, and Sandra C Fuchs. "Sleep duration and mortality in the elderly: a systematic review with meta-analysis". In: *BMJ Open* 6.2 (2016), e008119. DOI: 10.1136/bmjopen-2015-008119.

Simard, Patrice, Dave Steinkraus, and John Platt. "Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis". In: *International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2003. DOI: 10.1109/ICDAR.2003.1227801.

Simpson, Amber L. et al. "A large annotated medical image dataset for the development and evaluation of segmentation algorithms". In: *CoRR* abs/1902.09063 (2019).

Sobel, Irwin and Gary Feldman. "A 3×3 isotropic gradient operator for image processing". In: *Pattern Classification and Scene Analysis* (1973), pp. 271–272.

Song, Yeonsu, Terri Blackwell, Kristine Yaffe, Sonia Ancoli-Israel, Susan Redline, and Katie L. Stone. "Relationships between sleep stages and changes in cognitive function in older men: The MrOS sleep study". In: *Sleep* 38.3 (2015), pp. 411–421. ISSN: 1550-9109. DOI: 10.5665/sleep.4500.

Sørensen, T J. "A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons". In: *Biologiske Skrifter* 5.4 (1948). Kongelige Danske Videnskabernes Selskab, pp. 1–35.

Spira, Adam P., Terri Blackwell, Katie L. Stone, Susan Redline, Jane A. Cauley, Sonia Ancoli-Israel, and Kristine Yaffe. "Sleep-disordered breathing and cognition in older women". In: *Journal of the American Geriatrics Society* 56.1 (2008), pp. 45–50. ISSN: 0002-8614. DOI: 10.1111/j.1532-5415.2007.01506.x.

Stephansen, Jens B. et al. "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy". In: *Nature Communications* 9 (2018), p. 5229. DOI: 10.1038/s41467-018-07229-3.

Sudre, Carole H., Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. "Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations". In: *CoRR*. LNCS abs/1707.03237 (2017), pp. 240–248. DOI: 10.1007/978-3-319-67558-9_28.

Sun, Haoqi, Jian Jia, Balaji Goparaju, Guang-Bin Huang, Olga Sourina, Matt Travis Bianchi, and M Brandon Westover. "Large-Scale Automated Sleep Staging". In: *Sleep* 40.10 (Sept. 2017). zsx139. ISSN: 0161-8105. DOI: 10.1093/sleep/zsx139.

Supratak, Akara, Hao Dong, Chao Wu, and Yike Guo. "DeepSleepNet: a Model for Automatic Sleep Stage Scoring based on Raw Single-Channel EEG". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.11 (11 2017), pp. 1998–2008. DOI: 10.1109/TNSRE.2017.2721116.

Suri, Jasjit S and Rangaraj M Rangayyan. "Breast Imaging, Mammography, and Computer-Aided Diagnosis of Breast Cancer". In: SPIE Bellingham, WA, USA. 2006.

Szegedy, Christian, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. "Inception-v4, inception-resnet and the impact of residual connections on learning". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017. DOI: 10.1609/aaai.v31i1.11231.

Tack, Alexander, Anirban Mukhopadhyay, and Stefan Zachow. "Knee menisci segmentation using convolutional neural networks: Data from the Osteoarthritis Initiative". In: *Osteoarthritis Cartilage* 26 (2018), pp. 680–688. DOI: 10.1016/j.joca.2018.02.907.

Tan, Chuanqi, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. "A survey on deep transfer learning". In: *Artificial Neural Networks and Machine Learning–ICANN 2018:*

27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27. Springer. 2018, pp. 270–279. DOI: 10.1007/978-3-030-01424-7_27.

Team, R Core. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2019.

Terzano, Mario Giovanni et al. "Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep". In: *Sleep Medicine* 3.2 (2002), pp. 187–199. DOI: 10.1016/S1389-9457(02)00003-5.

Thorey, Valentin, Albert Bou Hernandez, Pierrick J. Arnal, and Emmanuel H. During. *AI vs Humans for the diagnosis of sleep apnea*. 2019. DOI: 10.1109/EMBC.2019.8856877.

Tobaldini, Eleonora, Elisa M Fiorelli, Monica Solbiati, Giorgio Costantino, Lino Nobili, and Nicola Montano. "Short sleep duration and cardiometabolic risk: from pathophysiology to clinical evidence". In: *Nature Reviews Cardiology* 16.4 (2018), pp. 213–224. DOI: 10.1038/s41569-018-0109-6.

Tsinalis, Orestis, Paul M. Matthews, and Yike Guo. "Automatic Sleep Stage Scoring Using Time-Frequency Analysis and Stacked Sparse Autoencoders". In: *Annals of Biomedical Engineering* 44.5 (2016), pp. 1587–1597. DOI: 10.1007/s10439-015-1444-y.

Vallat, Raphael and Matthew P Walker. "An open-source, high-performance tool for automated sleep staging". In: *Elife* 10 (2021), e70092. DOI: 10.7554/eLife.70092.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017).

Vilamala, Albert, Kristoffer Hougaard Madsen, and Lars Kai Hansen. "Deep Convolutional Neural Networks for Interpretable Analysis of EEG Sleep Stage Scoring". In: *CoRR* abs/1710.00633 (2017). DOI: 10.1109/MLSP.2017.8168133.

Virtanen, Pauli et al. "SciPy 1.0-Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

Vyazovskiy, Vladyslav, Umberto Olcese, Erin Hanlon, Yuval Nir, Chiara Cirelli, and Giulio Tononi. "Local sleep in awake rats". In: *Nature* 472.7344 (2011), pp. 443–447. DOI: 10.1038/nature10009.

Walker, H Kenneth, W Dallas Hall, and J Willis Hurst. "Clinical methods: the history, physical, and laboratory examinations". In: (1990).

Wang, Linwei, Qi Dou, P Thomas Fletcher, Stefanie Speidel, and Shuo Li. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I-VI*. Vol. 13431-13438. Springer Nature, 2022. DOI: 10.1007/978-3-031-16452-1.

Wang, Shuo, Mu Zhou, Zaiyi Liu, Zhenyu Liu, Dongsheng Gu, Yali Zang, Di Dong, Olivier Gevaert, and Jie Tian. "Central focused convolutional neural networks: Developing a data-driven model

for lung nodule segmentation". In: *Medical image analysis* 40 (2017), pp. 172–183. DOI: 10.1016/j.media.2017.06.014.

Warby, Simon, Sabrina Wendt, Peter Welinder, Emil GS Munk, Oscar Carrillo, Helge Bjarup Dissing Sørensen, Poul Jennum, Paul Peppard, Pietro Perona, and Emmanuel Mignot. "Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods". In: *Nature Methods* 11 (2014), pp. 385–392. DOI: 10.1038/nmeth.2855.

Weltens, Caroline, Johan Menten, Michel Feron, Erwin Bellon, Philippe Demaerel, Frederik Maes, Walter Van den Bogaert, and Emmanuel van der Schueren. "Interobserver variations in gross tumor volume delineation of brain tumors on computed tomography and impact of magnetic resonance imaging". In: *Radiotherapy and Oncology* 60.1 (2001), pp. 49–59. DOI: 10.1016/S0167-8140(01)00371-1.

Wernick, Miles N., Yongyi Yang, Jovan G. Brankov, Grigori Yourganov, and Stephen C. Strother. "Machine Learning in Medical Imaging". In: *IEEE Signal Processing Magazine* 27.4 (2010), pp. 25–38. DOI: 10.1109/MSP.2010.936730.

WHO et al. *Global spending on health: rising to the pandemic's challenges*. World Health Organization, 2022.

Wiegand, Thomas, Ramesh Krishnamurthy, Monique Kuglitsch, Naomi Lee, Sameer Pujari, Marcel Salathé, Markus Wenzel, and Shan Xu. "WHO and ITU establish benchmarking process for artificial intelligence in health". In: *The Lancet* 394.10192 (2019), pp. 9–11. ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(19)30762-7.

Wirth, Wolfgang, Felix Eckstein, Jochen Kemnitz, et al. "Accuracy and longitudinal reproducibility of quantitative femorotibial cartilage measures derived from automated U-Net-based segmentation of two different MRI contrasts: Data from the osteoarthritis initiative healthy reference cohort". In: *Magn Reson Mater Phys Biol Med* 34 (2021), pp. 337–354. DOI: 10.1007/s10334-020-00889-7.

Wittchen, H U et al. "The size and burden of mental disorders and other disorders of the brain in Europe 2010". In: *European Neuropsychopharmacology* 21.9 (2011), pp. 655–679. DOI: 10.1016/j.euroneuro.2011.07.018.

Wu, Eric, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E Ho, and James Zou. "How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals". In: *Nature Medicine* 27.4 (2021), pp. 582–584. DOI: 10.1038/s41591-021-01312-x.

Würfl, Tobias, Florin C Ghesu, Vincent Christlein, and Andreas Maier. "Deep learning computed tomography". In: *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part III 19*. Springer. 2016, pp. 432–440. DOI: 10.1007/978-3-319-46726-9_50.

Xia, Yingda, Fengze Liu, Dong Yang, Jinzheng Cai, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan Yuille, and Holger Roth. "3d semi-supervised learning with uncertainty-aware multi-view co-training". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020, pp. 3646–3655. DOI: 10.1109/WACV45572.2020.9093608.

Xu, Daozhang, Jente van der Voet, Nils Hansson, and et al. "Association between meniscal volume and development of knee osteoarthritis". In: *Rheumatology* 60 (2020), pp. 1392–1399. DOI: 10.1093/rheumatology/keaa522.

Younes, Magdy, Samuel T. Kuna, Allan I. Pack, James K. Walsh, Clete A. Kushida, Bethany Staley, and Grace W. Pien. "Reliability of the American Academy of Sleep Medicine Rules for Assessing Sleep Depth in Clinical Practice". In: *Journal of Clinical Sleep Medicine* 14.02 (2018), pp. 205–213. ISSN: 1550-9389, 1550-9397. DOI: 10.5664/jcsm.6934.

Younes, Magdy, Jill Raneri, and Patrick Hanly. "Staging Sleep in Polysomnograms: Analysis of Inter-Scorer Variability". In: *Journal of Clinical Sleep Medicine* 12.06 (2016), pp. 885–894. ISSN: 1550-9389, 1550-9397. DOI: 10.5664/jcsm.5894.

Young, Terry, Mari Palta, Jerome Dempsey, Paul E. Peppard, F. Javier Nieto, and K. Mae Hla. "Burden of sleep apnea: rationale, design, and major findings of the Wisconsin Sleep Cohort study". In: *WMJ: official publication of the State Medical Society of Wisconsin* 108.5 (2009), pp. 246–249. ISSN: 1098-1861.

Yu, Fisher and Vladlen Koltun. "Multi-Scale Context Aggregation by Dilated Convolutions". In: *International Conference on Learning Representations (ICLR)*. 2016. DOI: 10.48550/arXiv.1511.07122.

Zhang, Guo-Qiang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. "The National Sleep Research Resource: Towards a sleep data commons". In: *JAMIA* 25.10 (2018), pp. 1351–1358. ISSN: 1527-974X. DOI: 10.1093/jamia/ocy064.

Zhang, Xiaozhe, Xiaosong Dong, Jan W. Kantelhardt, Jing Li, Long Zhao, Carmen Garcia, Martin Glos, Thomas Penzel, and Fang Han. "Process and outcome for international reliability in sleep scoring". In: *Sleep and Breathing* 19.1 (2015), pp. 191–195. ISSN: 1520-9512, 1522-1709. DOI: 10.1007/s11325-014-0990-0.

Zhou, Zhaoye, Guang Zhao, Richard Kijowski, and Fang Liu. "Deep convolutional neural network for segmentation of knee joint anatomy". In: *Magn Reson Med* 80 (2018), pp. 2759–2770. DOI: 10.1002/mrm.27229.

Zoph, Barret and Quoc V Le. "Neural architecture search with reinforcement learning". In: *arXiv preprint arXiv:1611.01578* (2016). DOI: 10.48550/arXiv.1611.01578.

Appendices

Appendix A

Appendix for Paper A

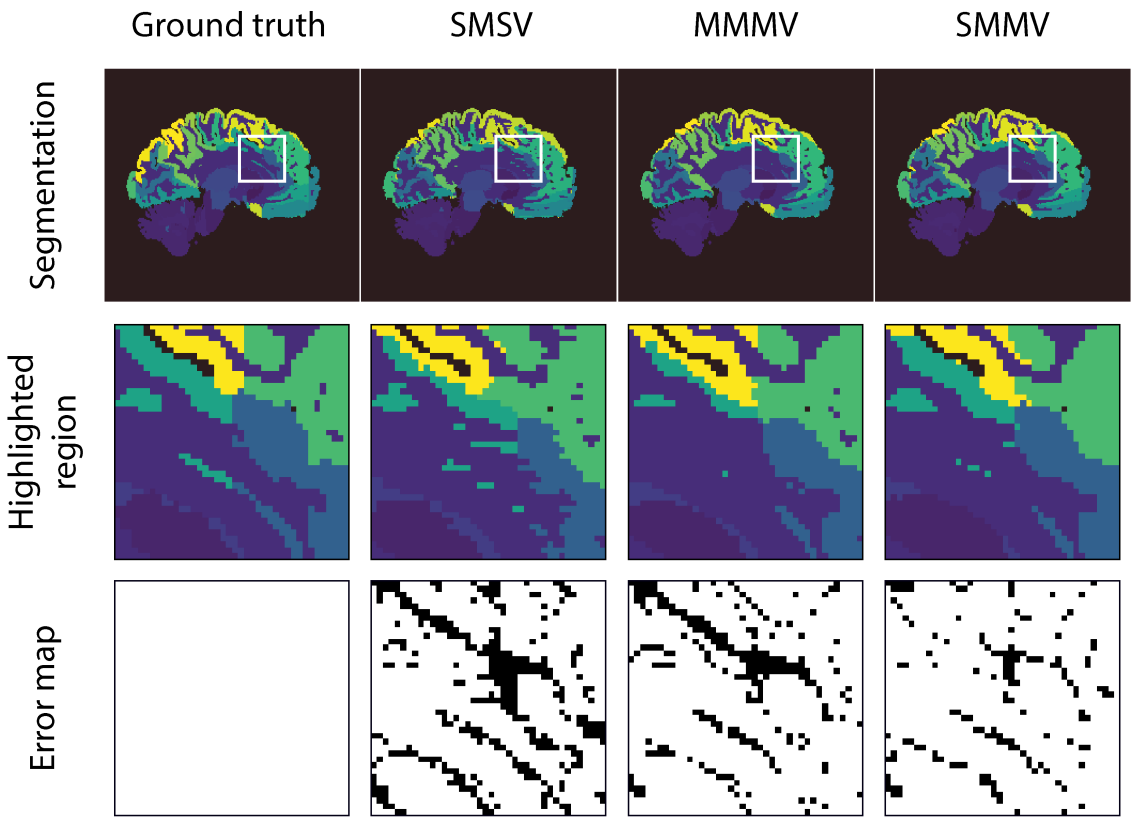


Figure A.1: Visual comparison of the typical performance improvements obtained on a random subject of the MICCAI dataset when going from a single U-Net model fit to a single plane (single-model-single-view, SMSV, second column) to an ensemble of such models (multi-model-multi-view, MMMV, third column) to the MPUnet (single-model-multi-view, SMMV, fourth column). The first row shows the full segmentation on a single 2D slice. The second row presents a zoom of the highlighted region shown in each image of row 1. The third row shows a binary error-map for the highlighted region with black pixels representing errors compared to the ground truth and white pixels representing correctly classified pixels.

Table A.1: Fixed hyperparameter set for the optimization of the MPUnet core model on any segmentation task.

Parameter	Value	Notes
Optimizer	Adam	The global learning rate is reduced by 10 % for every 2 consecutive epochs without validation performance improvements.
Learning rate -	$5 \cdot 10^{-5}$	
β_1 -	0.9	
β_2 -	0.999	
ϵ -	$1 \cdot 10^{-8}$	
Loss function	Cross entropy	
Regularization -	None	
Class balancing -	None	
Model Topology	2D U-Net	The input dimensions are inferred based on the sizes of the images of the training data cohort. The range of 128-512 is appropriate for typical compute systems, but may be expanded to work on larger images. Generalization properties outside of this suggested range have not been tested. Note that small images volumes may be oversampled.
Input dim -	128-512	
Depth -	4	
Up-sampling -	Nearest neighbour	
Activations -	ReLU	
Conv. kernel size -	3×3	
Max-pool kernel size -	2×2	
Padding -	True ('same')	
Batch normalization -	True	
Parameters -	$6.2 \cdot 10^7$	
Image sampling	Multi-Planar	Plane unit vectors are sampled uniformly from the 3-sphere with at least 60 deg angle between them.
Image interp -	Tri-linear	
Label interp -	Nearest-neighbour	
Num. planes -	6	
Non-linear aug.	RED*	Strength and smoothness sampled on-the-fly to produce variable deformations. *Random Elastic Deformations.
Strength, α -	uniform(100, 500)	
Elasticity, σ -	uniform(20, 30)	
Apply prob. -	1/3	
Loss weight -	1/3	
Pre-processing	Robust scaling	Image- and channel-wise scaling to (non-background) intensity distribution of median 0 and IQR 1.
Post-processing	None	
Batch size	8-16	16 by default, reduced by 2 until batches fit in GPU memory. A fraction of 1 minus the mean validation recall of a batch must contain non-background images (≥ 1 pixel of class $\neq 0$).
Foreground fraction -	1 - recall	
Training epochs	∞	Training continues until 15 consecutive epochs of without validation performance improvements.
Train images/epoch -	2500	
Val. images/epoch -	3500	
Early stopping criteria	Validation F1	Mean per-class F1 scores (excluding background) computed over all images of a validation epoch.
Model selection criteria	Validation F1	

Table A.2: F1 improvement on the MICCAI and MSD Task 4 datasets for a MPUnet of 2-9 planes relative to the mean performance of 9 single-plane models each fit to 1 of the 9 planes of the 9-plane MPUnet model. While the absolute performance benefit of using higher numbers of planes vary between the two tasks, the gains are monotonically increasing with views across both. Note that these results are only guiding as the experiments were conducted just once for each MPUnet.

Num. planes, $i =$	9	8	7	6	5	4	3	2
MICCAI	0.041	0.037	0.037	0.035	0.029	0.024	0.015	0.012
MSD T4	0.017	0.017	0.016	0.015	0.015	0.014	0.013	0.012

Table A.3: Mean F1 performance on the MICCAI dataset for MPUnets of $i \in \{3, 6, 9\}$ planes compared to ensembles of individual single-plane model each trained on a unique plane. Each single-plane model is optimized under the same set of hyperparameter as the MPUnet. Note that the single-planar ensembles have i times the parameters of their MPUnet counterparts divided evenly across its i sub-models.

Num. planes, $i =$	9	6	3
Single-Planar Ensemble	0.717 ± 0.019	0.714 ± 0.021	0.710 ± 0.024
Multi-Planar U-Net	0.743 ± 0.028	0.737 ± 0.027	0.717 ± 0.030

Table A.4: Detailed report of the MPUnet mean and standard deviation F1 (dice) performance on individual target classes across the 10 tasks of the Medical Segmentation Decathlon.

Dataset	Description	Class	F1 Score
Task 1	Brain Tumours	Edema	0.70 ± 0.20
		Non-enhancing tumor	0.43 ± 0.31
		Enhancing tumour	0.67 ± 0.22
Task 2	Cardiac	Left atrium	0.89 ± 0.09
Task 3	Liver & Tumour	Liver	0.94 ± 0.03
		Cancer	0.57 ± 0.32
Task 4	Hippocampus ROI.	Anterior	0.90 ± 0.03
		Posterior	0.88 ± 0.04
Task 5	Prostate	Peripheral zone	0.69 ± 0.13
		Transition zone	0.86 ± 0.07
Task 6	Lung Tumours	Cancer	0.59 ± 0.23
Task 7	Pancreas & Tumour	Pancreas	0.71 ± 0.14
		Cancer	0.25 ± 0.27
Task 8	Hepatic Ves. & Tumour	Vessel	0.59
		Tumour	0.38
Task 9	Spleen	Spleen	0.95
Task 10	Colon Cancer	Cancer primaries	0.28

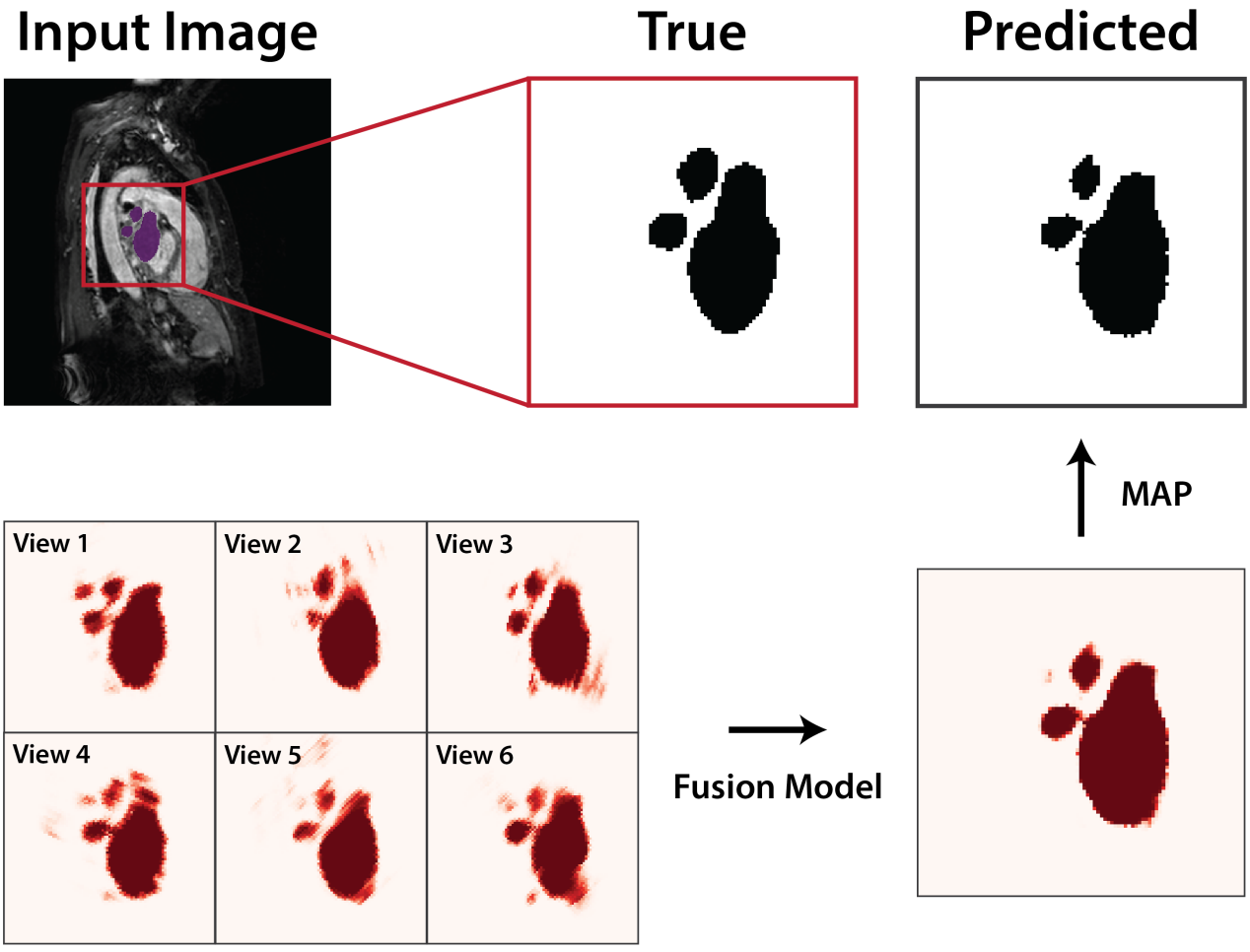


Figure A.2: Visualization of the benefit of the MPUNet test-time augmentation approach. A 2D slice from an input image is shown in the upper left panel with a highlighted region of interest to the right giving the ground truth (binary) label map for the left atrium of an image in the Medical Segmentation Decathlon Task 4 dataset. A single MPUnet predicts on the entire image volume along 6 planes and maps the predictions to the input image space, producing a set of 6 segmentation volumes. For each of those, the corresponding slice to the input image is shown in the lower left panel. Darker red colors indicate higher confidence of the model in the foreground class at the given pixel as seen in a given view. Note that while each confidence map matches the ground truth to a large extend, the model has both false positive and false negative confidence in certain areas of individual views. After passing the 6 segmentation maps through the fusion model (lower right), a much cleaner output is produced, which after thresholding (upper right) coresponds well to the ground truth.

Table A.5: Comparison of the Multi-Planar UNet and a 3D UNet of identical topology (all 2D operations replaced by 3D operations) on the three non-challenge benchmark datasets MICCAI, HaRP and OAI as well as the Medical Segmentation Decathlon (MSD) Task 4 dataset (hippocampus in region-of-interest). The two models were trained under identical optimization parameters. The shown scores are mean per-class F1 scores pooled across three separate training and evaluation sessions. The MSD Task 4 dataset experiments were conducted on random splits of the challenge training data, as we do not have access to the test set. The 3D UNet was trained on isotropic ROIs of 64-cube voxels with random rotations and 3D random elastic deformations applied at batch-sampling time. This was done to emulate the benefit of the MPUNet's significant data augmentation. The sampled voxel-resolution was identical to that chosen for the MPUNet. The 3D model has a total of 90 million parameters against the 62 of the MPUnet. The MSD Task 4 dataset consists of small cut-out regions of interest spanning narrowly around the hippocampus to segment, and was include here to study the performance of the 3D model when the entire input image fits within the 64-cube input patch. **Note:** The OAI dataset used for those experiments was a smaller subset of the full dataset for which results are displayed in Table 5.1 (no follow-up scans included, specifically).

	MICCAI	HaRP	OAI	MSD T4
3D U-Net w. rotations	0.74 ± 0.04	0.84 ± 0.05	0.81 ± 0.07	0.87 ± 0.04
Multi-Planar U-Net	0.74 ± 0.03	0.85 ± 0.03	0.84 ± 0.07	0.88 ± 0.04

Table A.6: MPUnet base model topology (U-Net type) for images sampled with pixel dim $q = 256$. Note: Convolution strides of 1×1 where used in all layers.

Layer name	**Output dim**	**Kernel dim**	**Filters**	**Activation**	**Pad**
Input	$256 \times 256 \times C$	-	-	-	-
conv_1_1	$256 \times 256 \times 90$	3×3	90	ReLU	same
conv_1_2	$256 \times 256 \times 90$	3×3	90	ReLU	same
bn_1	$256 \times 256 \times 90$	-	-	-	-
pool_1	$128 \times 128 \times 90$	2×2	-	-	valid
conv_2_1	$128 \times 128 \times 181$	3×3	181	ReLU	same
conv_2_2	$128 \times 128 \times 181$	3×3	181	ReLU	same
bn_2	$128 \times 128 \times 181$	-	-	-	-
pool_2	$64 \times 64 \times 181$	2×2	-	-	valid
conv_3_1	$64 \times 64 \times 362$	3×3	362	ReLU	same
conv_3_2	$64 \times 64 \times 362$	3×3	362	ReLU	same
bn_3	$64 \times 64 \times 362$	-	-	-	-
pool_3	$32 \times 32 \times 362$	2×2	-	-	valid
conv_4_1	$32 \times 32 \times 724$	3×3	724	ReLU	same
conv_4_2	$32 \times 32 \times 724$	3×3	724	ReLU	same
bn_4	$32 \times 32 \times 724$	-	-	-	-
pool_4	$16 \times 16 \times 724$	2×2	-	-	valid
conv_5_1	$16 \times 16 \times 1448$	3×3	1448	ReLU	same
conv_5_2	$16 \times 16 \times 1448$	3×3	1448	ReLU	same
up_1	$32 \times 32 \times 1448$	2×2	-	-	-
conv_6_0	$32 \times 32 \times 724$	2×2	724	ReLU	same
bn_6	$32 \times 32 \times 724$	-	-	-	-
merge(bn4, bn6)	$32 \times 32 \times 1448$	-	-	-	-
conv_6_1	$32 \times 32 \times 724$	3×3	724	ReLU	same
conv_6_2	$32 \times 32 \times 724$	3×3	724	ReLU	same
bn_7	$32 \times 32 \times 724$	-	-	-	-
up_2	$64 \times 64 \times 724$	2×2	-	-	-
conv_7_0	$64 \times 64 \times 362$	2×2	362	ReLU	same
bn_8	$64 \times 64 \times 362$	-	-	-	-
merge(bn3, bn8)	$64 \times 64 \times 724$	-	-	-	-
conv_7_1	$64 \times 64 \times 362$	3×3	362	ReLU	same
conv_7_2	$64 \times 64 \times 362$	3×3	362	ReLU	same
bn_9	$64 \times 64 \times 362$	-	-	-	-
up_3	$128 \times 128 \times 362$	2×2	-	-	-
conv_8_0	$128 \times 128 \times 181$	2×2	181	ReLU	same
bn_10	$128 \times 128 \times 181$	-	-	-	-
merge(bn2, bn10)	$128 \times 128 \times 362$	-	-	-	-
conv_8_1	$128 \times 128 \times 181$	3×3	181	ReLU	same
conv_8_2	$128 \times 128 \times 181$	3×3	181	ReLU	same
bn_11	$128 \times 128 \times 181$	-	-	-	-
up_4	$256 \times 256 \times 181$	2×2	-	-	-
conv_9_0	$256 \times 256 \times 90$	2×2	90	ReLU	same
bn_12	$256 \times 256 \times 90$	-	-	-	-
merge(bn1, bn12)	$256 \times 256 \times 180$	-	-	-	-
conv_9_1	$256 \times 256 \times 90$	3×3	90	ReLU	same
conv_9_2	$256 \times 256 \times 90$	3×3	90	ReLU	same
bn_13	$256 \times 256 \times 90$	-	-	-	-
output	$256 \times 256 \times K$	1×1	K	softmax	-

Trainable parameters: $62,062,342$ **(for** $K = 135$**,** $C = 1$**)**

Appendix B

Appendix for Paper B

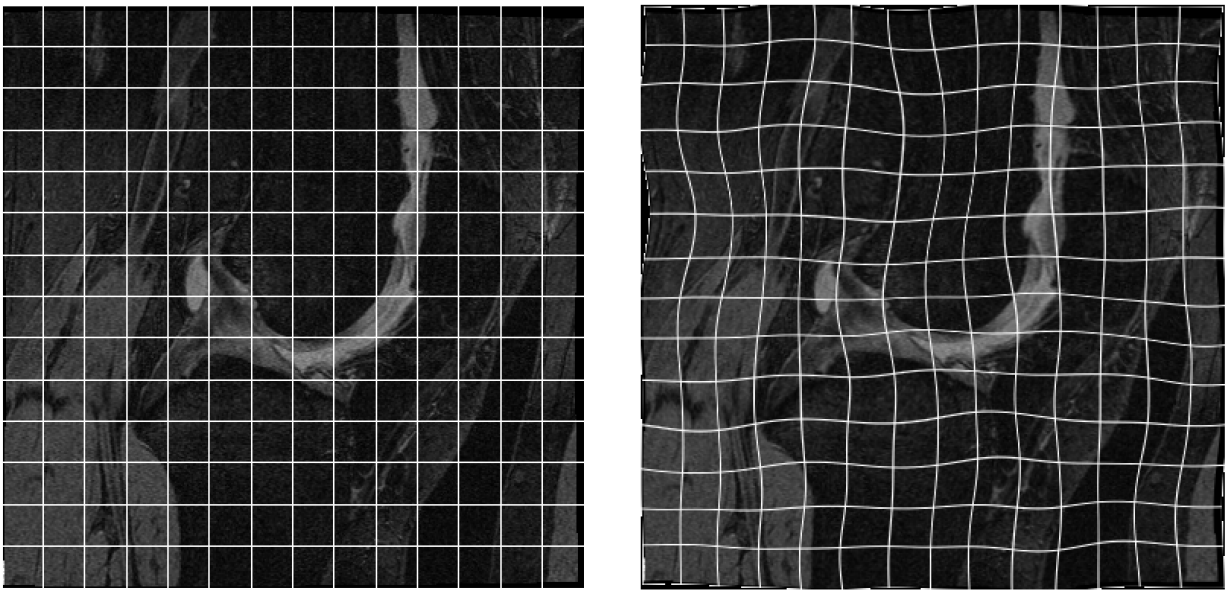


Figure B.1: Visualization of the effect of random elastic deformations. (a) Input image. (b) Augmented image.

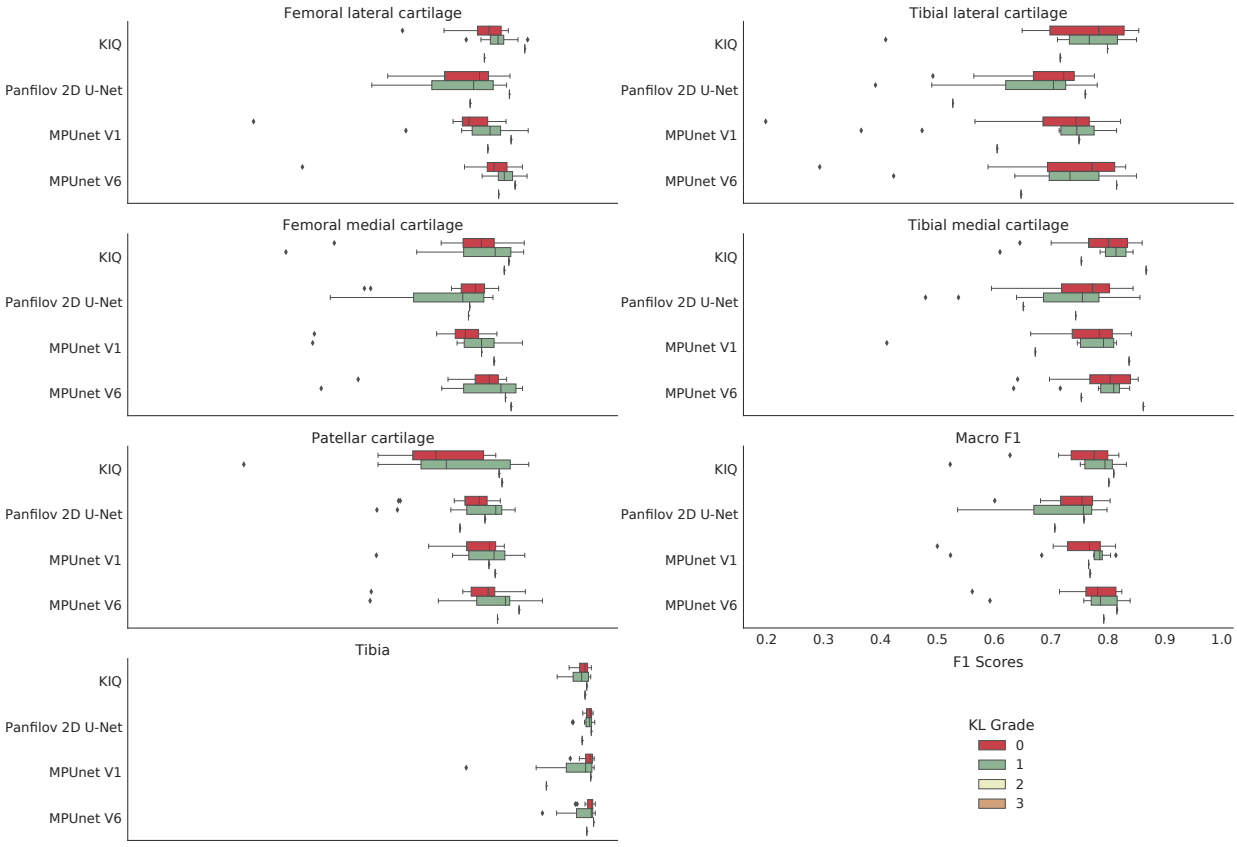


Figure B.2: Box-plots showing the distribution of dice scores for the MPUnet, KIQ and the Panfilov 2D U-Net on the PROOF dataset grouped according to the KL-grade score of the individual MRIs.

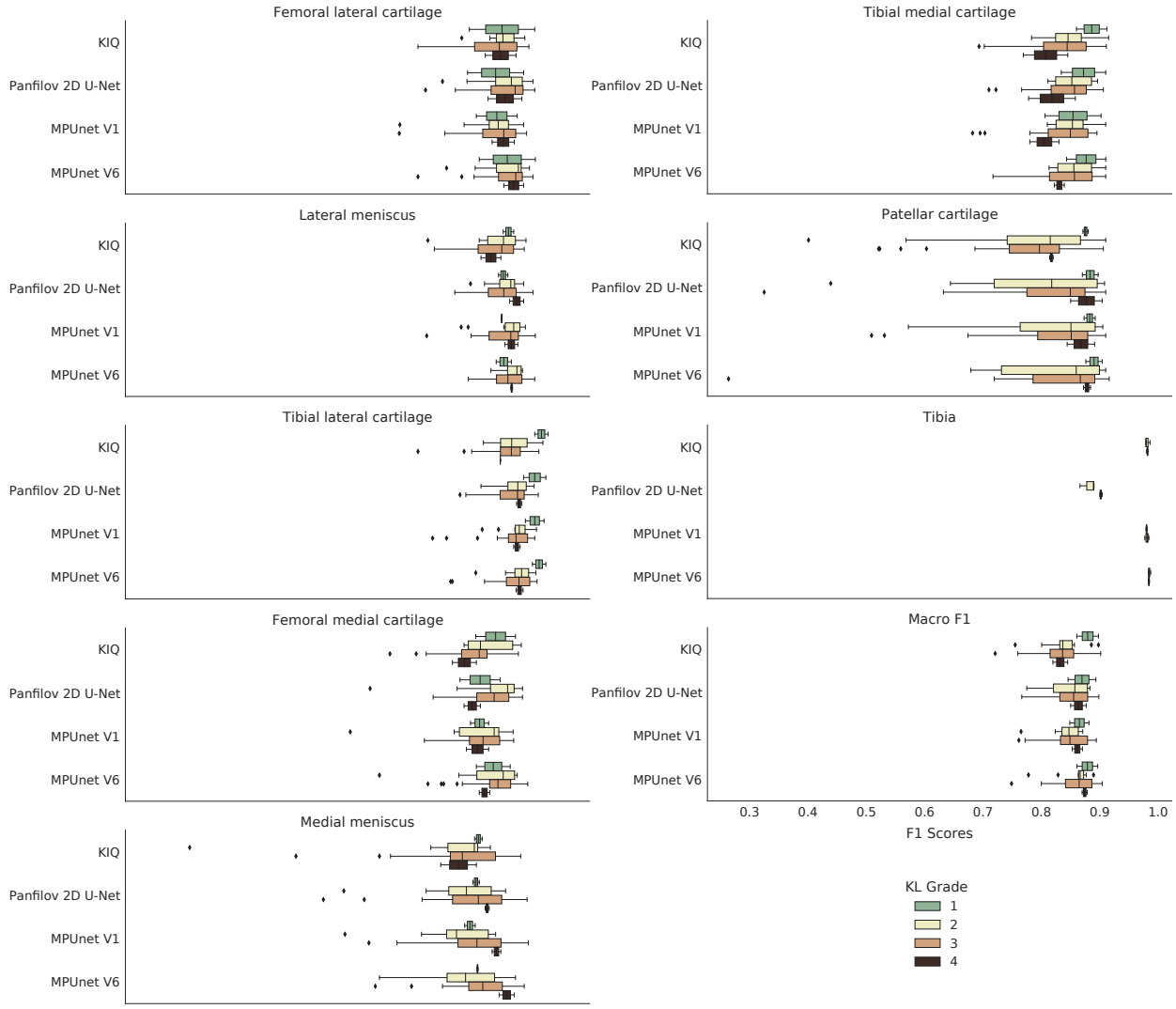


Figure B.3: Box-plots showing the distribution of dice scores for the MPUnet, KIQ and the Panfilov 2D U-Net on the OAI dataset grouped according to the KL-grade score of the individual MRIs.

Appendix C

Appendix for Paper C

Table C.1: Brief characterization of typical features of the 5 sleep stages as defined by the AASM manual (Iber et al. 2007).

Name	Encoding	Description
Wake	W	Spans wakefulness to drowsiness. Consists of at least 50% alpha waves (8-13 Hz EEG signals). Rapid and reading eye movements. Eye blinks may occur.
Non-REM 1	N1	Short, light sleep stage comprising about 5%-10% of a night's sleep. Dominated by theta waves (4-7 Hz EEG signals). Slow eye movements in W $\rightarrow$ N1 transition. Some EMG activity, but lower than wake.
Non-REM 2	N2	Comprises 40%-50% of a normal night's sleep. EEG dispalys theta-waves like N1, but intercepted by so-called K-complexes and/or sleep spindles (short bursts of 13-16Hz EEG signal).
Non-REM 3	N3	Comprises about 20%-25% of a typical night's sleep. High amplitude, slow 0.3-3 Hz EEG signals. Low EMG activity.
REM	R	Rapid-eye-movements may occur. Displays both theta waves and alpha (like wake), but typically 1-2 Hz slower. EMG significantly reduced. Dreaming may occur this stage, which comprises 20%-25% of the night's sleep.

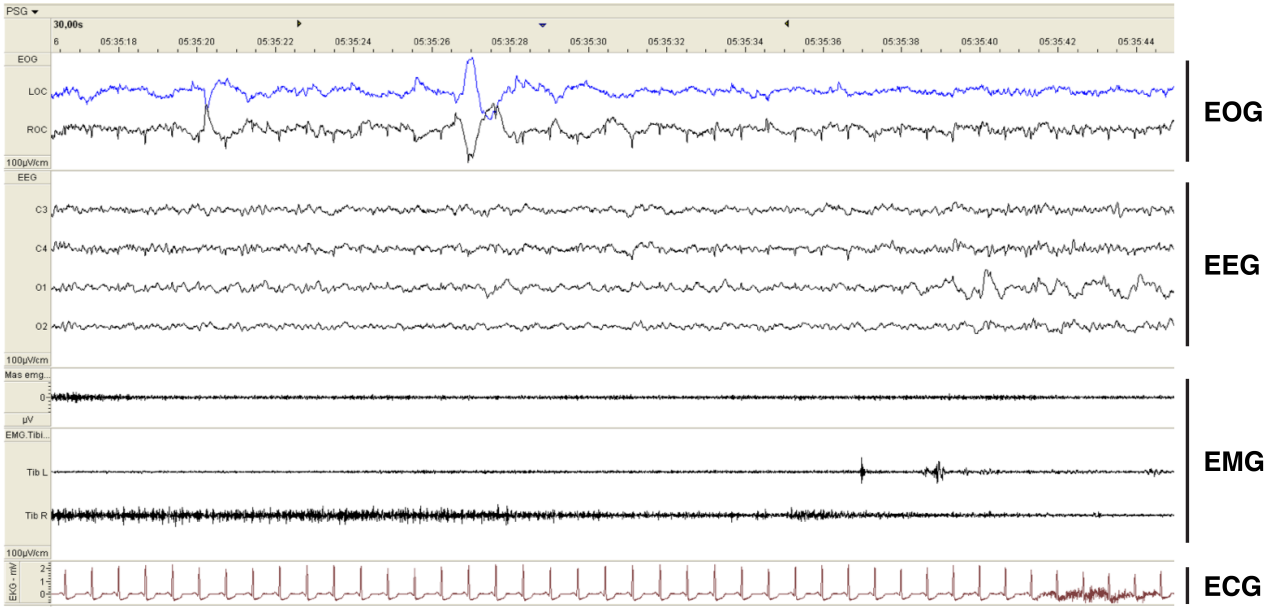


Figure C.1: A segment of 30 seconds of a typical polysomnography (PSG) study showing multiple EOG, EEG, EMG and ECG channels. Human experts evaluate segments such as this and assign it to one of the sleep stages in {W, N1, N2, N3, R}. In most experiments of this study, U-Time considers only a single EEG channel (for instance C3, as seen above).

Table C.2: U-Time model topology. Layer dimensions below are valid for $i = 3000$, $C = 1$, $T = 35$, $K = 5$. BN = batch normalization. All convolution kernels in layer 1 to 16 (the encoder) are dilated to with 9.

ID	Layer Type	Output dim	Kernel	Filters	Activation	Pad
1	Input	$35 \times 3000 \times 1$	-	-	-	-
2	Reshape	105000×1	-	-	-	-
3	Convolution → BN	105000×16	5	16	ReLU	same
4	Convolution → BN	105000×16	5	16	ReLU	same
5	Max Pool	10500×16	10	-	-	valid
6	Convolution → BN	10500×32	5	32	ReLU	same
7	Convolution → BN	10500×32	5	32	ReLU	same
8	Max Pool	1312×32	8	-	-	valid
9	Convolution → BN	1312×64	5	64	ReLU	same
10	Convolution → BN	1312×64	5	64	ReLU	same
11	Max Pool	218×64	6	-	-	valid
12	Convolution → BN	218×128	5	128	ReLU	same
13	Convolution → BN	218×128	5	128	ReLU	same
14	Max Pool	54×128	4	-	-	valid
15	Convolution → BN	54×256	5	256	ReLU	same
16	Convolution → BN	54×256	5	256	ReLU	same
17	Up-sample	216×256	4	-	-	-
18	Convolution → BN	216×128	4	128	ReLU	same
19	Crop & Concat(13, 18)	216×256	-	-	-	-
20	Convolution → BN	216×128	5	128	ReLU	same
21	Convolution → BN	216×128	5	128	ReLU	same
22	Up-sample	1296×128	6	-	-	-
23	Convolution → BN	1296×64	6	64	ReLU	same
24	Crop & Concat(10, 23)	1296×128	-	-	-	-
25	Convolution → BN	1296×64	5	64	ReLU	same
26	Convolution → BN	1296×64	5	64	ReLU	same
27	Up-sample	10368×64	8	-	-	-
28	Convolution → BN	10368×32	8	32	ReLU	same
29	Crop & Concat(7, 28)	10368×64	-	-	-	-
30	Convolution → BN	10368×32	5	32	ReLU	same
31	Convolution → BN	10368×32	5	32	ReLU	same
32	Up-sample	103680×32	10	-	-	-
33	Convolution → BN	103680×16	10	16	ReLU	same
34	Crop & Concat(4, 33)	103680×32	-	-	-	-
36	Convolution → BN	103680×16	5	16	ReLU	same
35	Convolution → BN	103680×16	5	16	ReLU	same
36	Convolution	103680×5	1	5	TanH	same
37	Zero padding	105000×5	-	-	-	-
38	Reshape	$35 \times 3000 \times 5$	-	-	-	-
38	Average Pooling	35×5	-	-	-	valid
39	Convolution	35×5	1	5	Softmax	same

Trainable parameters: $1,187,589$

Table C.3: Hyperparameters used for all datasets.

Parameter	Value	Notes
Optimizer	Adam	We employ a fixed learning rate across all datasets. See Kingma et al. (2015).
Learning rate -	$5 \cdot 10^{-6}$	
β_1 -	0.9	
β_2 -	0.999	
ϵ -	$1 \cdot 10^{-8}$	
Loss function	Dice loss	See Crum et al. (2006) and Sudre et al. (2017).
Regularization -	None	
Class balancing -	Uniform (None)	
Base Topology	1D U-Net	The input dimensionality is the number of data points in a single PSG segment (one segment is 30 seconds in typical sleep staging, giving input dimensionality 3000 for sample rate $S = 100$). T is the number of contiguous segments the model operates on at once. T may be dynamically adjusted. Cropping and zero-padding is needed to decode to dimensions equal to the input, see Ioffe et al. (2015), Odena et al. (2016), Ronneberger et al. (2015), and Yu et al. (2016).
Input dim -	3000	
Window size (T) -	35	
Depth -	4	
Up-sampling -	Nearest neighbour	
Activations -	ReLU	
Conv. kernel size -	5	
Conv. kernel dilation size -	9	
Max-pool kernel size -	$\{10, 8, 6, 4\}$	
Padding -	True ('same')	
Batch normalization -	True	
Parameters -	$\approx 1.2 \cdot 10^6$	
Pre-processing	Robust scaling	Record- and channel-wise transformation to distribution of median 0 and IQR 1.
Post-processing	None	
Re-sampling (S)	100 Hz	Re-sampling uses polyphase filtering (implementation: `scipy.signal.resample_poly`, see Virtanen et al. 2020).
Batch size (B)	12	For each member of a batch, a class from the label set {W, N1, N2, N3, R} is determined by uniform sampling. A random PSG record that contains the given class is sampled, from which the input window is sampled randomly so that the selected class is present somewhere in the window.
Class sampling prob. -	Uniform	
Training epochs	∞	Training continues until 150 consecutive epochs without validation performance improvements. L is the number of 30 second segments in the dataset.
Steps per epoch	$\lceil L/T/B \rceil$	
Early stopping criteria	Validation F1	Mean per-class F1 scores (excluding background) computed over all images of a validation epoch.
Model selection criteria	Validation F1	

Table C.4: U-Time per-record results. Values shown are F1/dice scores computed across all PSG records in each dataset. Each cell displays the mean F1 ± 1 standard deviation, with the lowest and highest observed F1 score across the records given in the line below indicated by ↓ and ↑ respectively.

Dataset	W	N1	N2	N3	REM
S-EDF-39	0.87 ± 0.13 $\downarrow 0.34 \uparrow 0.99$	0.49 ± 0.16 $\downarrow 0.05 \uparrow 0.81$	0.85 ± 0.11 $\downarrow 0.25 \uparrow 0.94$	0.81 ± 0.17 $\downarrow 0.12 \uparrow 0.96$	0.83 ± 0.16 $\downarrow 0.04 \uparrow 0.97$
S-EDF-153	0.89 ± 0.08 $\downarrow 0.55 \uparrow 0.99$	0.51 ± 0.13 $\downarrow 0.04 \uparrow 0.76$	0.83 ± 0.09 $\downarrow 0.40 \uparrow 0.96$	0.57 ± 0.30 $\downarrow 0.00 \uparrow 1.00$	0.79 ± 0.16 $\downarrow 0.00 \uparrow 0.98$
Physio-18	0.78 ± 0.16 $\downarrow 0.00 \uparrow 0.99$	0.57 ± 0.14 $\downarrow 0.00 \uparrow 0.87$	0.81 ± 0.12 $\downarrow 0.00 \uparrow 0.98$	0.69 ± 0.27 $\downarrow 0.00 \uparrow 1.00$	0.78 ± 0.23 $\downarrow 0.00 \uparrow 1.00$
DCSM	0.97 ± 0.04 $\downarrow 0.67 \uparrow 1.00$	0.47 ± 0.13 $\downarrow 0.00 \uparrow 0.80$	0.83 ± 0.11 $\downarrow 0.29 \uparrow 0.96$	0.76 ± 0.24 $\downarrow 0.00 \uparrow 0.97$	0.80 ± 0.20 $\downarrow 0.00 \uparrow 0.98$
ISRUC	0.84 ± 0.11 $\downarrow 0.42 \uparrow 0.97$	0.53 ± 0.12 $\downarrow 0.11 \uparrow 0.73$	0.77 ± 0.12 $\downarrow 0.07 \uparrow 0.92$	0.86 ± 0.10 $\downarrow 0.42 \uparrow 0.99$	0.73 ± 0.22 $\downarrow 0.00 \uparrow 0.99$
CAP	0.70 ± 0.22 $\downarrow 0.00 \uparrow 0.99$	0.28 ± 0.16 $\downarrow 0.00 \uparrow 0.66$	0.74 ± 0.14 $\downarrow 0.30 \uparrow 0.93$	0.79 ± 0.15 $\downarrow 0.10 \uparrow 0.95$	0.73 ± 0.23 $\downarrow 0.00 \uparrow 0.95$
SVUH-UCD	0.73 ± 0.12 $\downarrow 0.52 \uparrow 0.92$	0.46 ± 0.12 $\downarrow 0.28 \uparrow 0.66$	0.75 ± 0.16 $\downarrow 0.25 \uparrow 0.95$	0.79 ± 0.21 $\downarrow 0.00 \uparrow 0.98$	0.67 ± 0.26 $\downarrow 0.04 \uparrow 0.93$

Table C.5: U-Time ($C = 1$) confusion matrix for dataset Sleep-EDF-39

	Wake	N1	N2	N3	REM
Wake	**6980**	740	244	22	260
N1	205	**1624**	604	15	356
N2	360	615	**15182**	982	660
N3	25	7	777	**4892**	2
REM	204	516	523	0	**6474**

Table C.6: U-Time ($C = 1$) confusion matrix for dataset Sleep-EDF-153

	Wake	N1	N2	N3	REM
Wake	**58676**	5650	650	40	790
N1	2364	**12067**	5172	132	1787
N2	335	5478	**57437**	3491	2391
N3	10	69	2974	**9978**	8
REM	323	2510	2280	83	**20639**

Table C.7: U-Time ($C = 1$) confusion matrix for dataset Physionet-2018

	Wake	N1	N2	N3	REM
Wake	**133594**	20295	2473	96	1487
N1	22006	**83149**	22744	183	8896
N2	6834	32279	**304191**	25593	8924
N3	493	214	17779	**84006**	100
REM	3165	9095	6782	138	**97684**

Table C.8: U-Time ($C = 1$) confusion matrix for dataset DCSM

	Wake	N1	N2	N3	REM
Wake	**341590**	5681	2326	316	4396
N1	2839	**11128**	4804	19	2350
N2	1888	6037	**94237**	6586	4279
N3	195	33	7156	**36200**	53
REM	1931	1733	2522	435	**40205**

Table C.9: U-Time ($C = 1$) confusion matrix for dataset ISRUC

	Wake	N1	N2	N3	REM
Wake	**17237**	1892	512	33	751
N1	1349	**6505**	2316	66	1254
N2	359	2649	**22135**	1878	1174
N3	38	10	2332	**14876**	26
REM	363	974	938	56	**9589**

Table C.10: U-Time ($C = 1$) confusion matrix for dataset CAP

	Wake	N1	N2	N3	REM
Wake	**14126**	1532	1779	411	1004
N1	1149	**1412**	997	84	797
N2	1244	1351	**28629**	3477	2195
N3	135	32	4560	**19069**	296
REM	760	870	2187	394	**13429**

Table C.11: U-Time ($C = 1$) confusion matrix for dataset SVUH-UCD

	Wake	N1	N2	N3	REM
Wake	**3537**	739	227	18	186
N1	783	**1704**	525	8	383
N2	174	601	**5423**	410	377
N3	9	7	310	**2328**	9
REM	207	300	212	22	**2275**

Table C.12: U-Time multi-channel results across 4 datasets. Dataset sizes and evaluation types match those of Table 8.2 in the main text. Specefic channels used: Sleep-EDF-153: EEG Fpz-Cz, EMG submental, EOG horizontal. Physionet-2018: EEG C3-M2, EEG O1-M2, EMG CHEST. DCSM: EEG C3-M2, EOG E2-M2. ISRUC: EEG C3-M2, EOG ROC-M1.

		Global F1 scores					
Dataset	Channels	W	N1	N2	N3	REM	mean
S-EDF-153	EEG + EMG + EOG	0.92	0.51	0.82	0.72	0.84	0.76
Physio-18	2×EEG + EMG	0.83	0.58	0.83	0.79	0.83	0.77
DCSM	EEG + EOG	0.97	0.51	0.83	0.83	0.86	0.80
ISRUC	EEG + EOG	0.88	0.55	0.79	0.87	0.83	0.78

Table C.13: Hyperparameter experiments for our re-implemented DeepSleepNet (Supratak et al. 2017) on the Sleep-EDF-39 dataset. The 5-CV hyperparameter experiments were conducted on 25 records only in order to speed up computation. Thus, the performance scores should not be compared directly to the paper re-implementation results (which are based on all 39 records in a 20-CV evaluation), but rather to the *baseline* experiment.

		Global F1 scores					
Experiment	Eval.	W	N1	N2	N3	REM	mean
Paper re-implementation	20-CV	0.86	0.41	0.87	0.83	0.81	0.76
Baseline	5-CV	0.85	0.39	0.86	0.89	0.79	0.76
Smaller CNN filters	5-CV	0.84	0.31	0.86	0.87	0.77	0.73
Larger CNN filters	5-CV	0.84	0.30	0.87	0.87	0.76	0.73
Two CNN layers	5-CV	0.84	0.26	0.85	0.85	0.76	0.71
Four CNN layers	5-CV	0.84	0.31	0.86	0.89	0.78	0.74
One RNN layer	5-CV	0.85	0.35	0.86	0.88	0.78	0.74
Three RNN layers	5-CV	0.76	0.41	0.85	0.85	0.75	0.72
Short sequences ($T = 10$)	5-CV	0.83	0.31	0.86	0.87	0.75	0.72
Long sequences ($T = 50$)	5-CV	0.83	0.34	0.86	0.86	0.74	0.73
LSTM $\rightarrow$ GRU	5-CV	0.84	0.35	0.86	0.87	0.74	0.73
LSTM 64 cells	5-CV	0.85	0.32	0.85	0.84	0.78	0.73
LSTM 256 cells	5-CV	0.85	0.31	0.86	0.86	0.77	0.73
Dropout $\rightarrow$ Zoneout (5%)	5-CV	0.80	0.34	0.85	0.88	0.77	0.73
Dropout $\rightarrow$ Zoneout (10%)	5-CV	0.80	0.39	0.84	0.87	0.77	0.73
CNN filter size 3-ensemble	5-CV	0.86	0.34	0.87	0.88	0.79	0.75
FPZ+CZ+EOG ensemble	5-CV	0.91	0.40	0.89	0.85	0.87	0.78

Table C.14: Hyperparameter experiments for our re-implemented DeepSleepNet (Supratak et al. 2017) on the DCSM dataset. The hyperparameter experiments were conducted on a 100-records subset of the DCSM dataset to speed up computation. Thus, performance scores should not be compared to the results in Table 8.2 directly, but rather to the *baseline* experiment.

		Global F1 scores					
Experiment	Eval.	W	N1	N2	N3	REM	mean
Baseline	5-CV	0.95	0.33	0.81	0.77	0.80	0.73
Smaller CNN filters	5-CV	0.95	0.37	0.80	0.79	0.81	0.74
Larger CNN filters	5-CV	0.94	0.34	0.81	0.77	0.80	0.73
LSTM $\rightarrow$ GRU	5-CV	0.95	0.33	0.80	0.76	0.80	0.73
Short sequences ($T = 10$)	5-CV	0.95	0.32	0.80	0.75	0.78	0.72
Long sequences ($T = 50$)	5-CV	0.95	0.32	0.79	0.78	0.79	0.73
Four input signals ensemble	5-CV	0.96	0.36	0.83	0.80	0.81	0.75

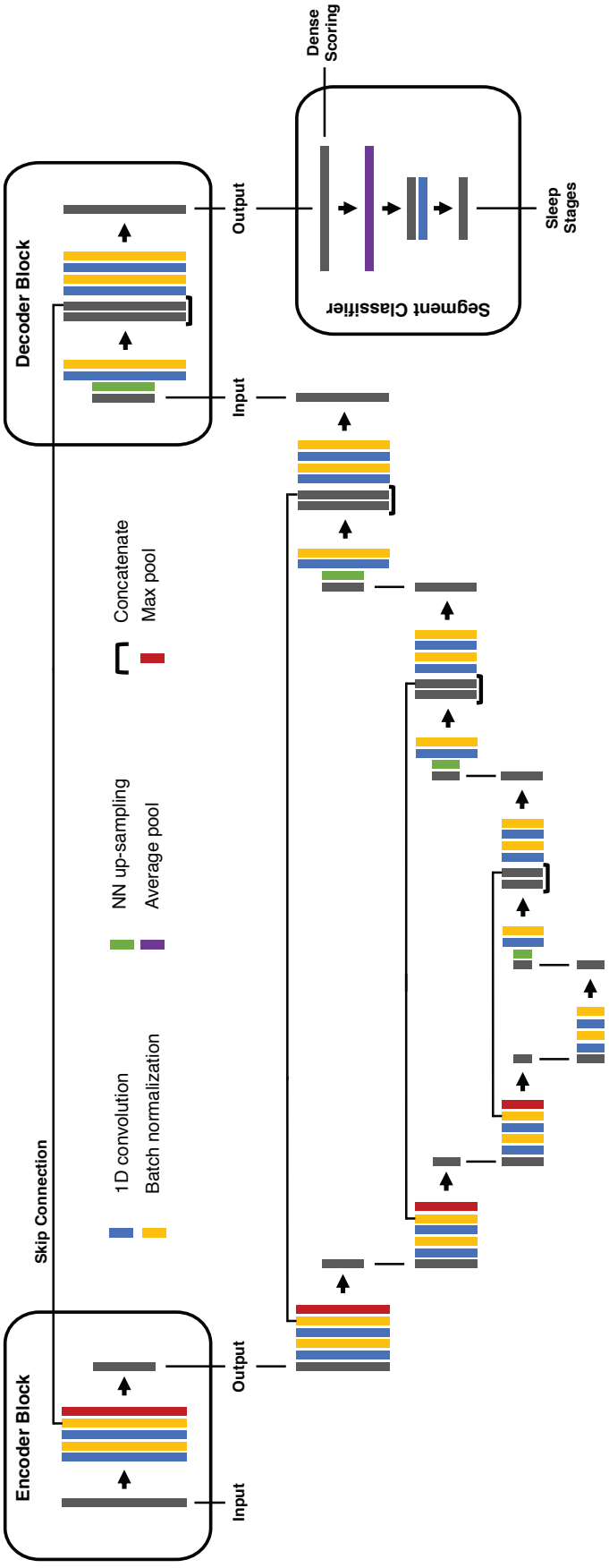


Figure C.2: Expanded structural overview of the U-Time architecture.

Appendix D

Appendix for Paper D

D.1 Supplementary Note: Datasets

In the following, we briefly describe the datasets considered in this study.

ABC The Apnea, Bariatric surgery, and CPAP (ABC) study consists of PSG recordings from patients with severe obstructive sleep apnea (OSA) and morbid obesity (BMI of 35-45) (Bakker et al. 2018; G.-Q. Zhang et al. 2018). The study addressed the effect of bariatric (weight loss) surgery in comparison to continuous positive airway pressure (CPAP) therapy for the treatment of OSA. The study pooled data from two different US sleep programs and spans a demographically diverse group of OSA patients. 53 subjects were enrolled in the original study, of which 49 were available to us for our work. EEG and EOG signals were recorded at 256 Hz and hardware low-pass filtered at 105 Hz and high-pass filtered at 0.16 Hz. Hypnograms were scored according to the AASM criteria. For more information, we refer to `https://doi.org/10.25822/nx52-bc11` and `https://clinicaltrials.gov/ct2/show/NCT01187771`.

CCSHS The Cleveland Children's Sleep and Health Study (CCSHS) is a large cohort of children and adolescents originally studied at ages 8-11 (Rosen, Larkin, et al. 2003; G.-Q. Zhang et al. 2018). The cohort is a stratified random sample of full-term and pre-term children born at 3 different hospitals around Cleveland, Ohio, US between 1988 and 1993. We considered PSG data as recorded in-lab during the third and final longitudinal visit which took place between 2006 and 2010. In our study, we had access to 515 samples of adolescents aged 16-19. EEG and EOG signals were recorded at 128 Hz and hardware high-pass filtered at 0.15 Hz. Hypnograms were scored according to the AASM criteria. For more information, we refer to `https://doi.org/10.25822/cg2n-4y91`.

CFS The Cleveland Family Study (CFS) is a large, family-based study of sleep apnea consiting of 2284 subjects from 361 families studied longitudinally between 1990 and 2006 (Redline, Tishler, et al. 1995; G.-Q. Zhang et al. 2018). We considered data from the last visit (2006) at which full overnight PSG were measured. 730 subjects from 144 families participated in this study. When splitting data from `CFS` into train/test splits, we ensured that all family members appear in the same split. EEG and EOG signals were recorded at 128 Hz and hardware low-pass filtered at 105 Hz and high-pass filtered at 0.16 Hz. Hypnograms were scored according to the AASM criteria. For more information

on this dataset, we refer to https://doi.org/10.25822/jmyx-mz90.

CHAT The Childhood Adenotonsillectomy Trial (CHAT) studied the effect of adenotonsillectomy surgery (removal of tonsils and adenoids) on mild to moderate obstructive sleep apnea (OSA) in children ages 5-10 years (C. L. Marcus et al. 2013; Redline, Amin, et al. 2011; G.-Q. Zhang et al. 2018). Subjects were assessed with full PSG at baseline and after a 7-month period. Study participants were recruited from 6 US sleep centres in Massachusetts, Missouri, New York, Ohio and Pennsylvania. We considered a total of 1638 PSG records from 1232 subjects (452 baseline, 407 follow-up, 779 control). Record chat-baseline-300927 was excluded due to missing EOG channels. EEG and EOG signals were recorded at 200 Hz or higher (varies between studies) with variable hardware filtering applied depending on the acquisition system. Hypnograms were scored according to the AASM criteria. For more information, we refer to https://doi.org/10.25822/d68d-8g03 and https://clinicaltrials.gov/ct2/show/NCT00560859.

HPAP The Home Positive Airway Pressure (HomePAP, abbreviated HPAP in this study) was a multi-site study with patients enrolled from 7 different US academic sleep centres to study the effectiveness of home-based portable monitoring as compared to laboratory-based PSG for the diagnosis and treatment of obstructive sleep apnea (OSA) in adults at least 18 years of age (Rosen, Auckley, et al. 2012; G.-Q. Zhang et al. 2018). The study included 373 subjects of which we consider only the 247 who underwent lab-based PSG recordings. We excluded 9 subjects (IDs 1600052, 1600138, 1600280, 1600047, 1600194, 1600361, 1600087, 1600368, and 1600203) due to missing EOG and/or reference channels. EEG and EOG signals were recorded at 200 Hz with no hardware filtering applied. Hypnograms were scored according to the AASM criteria. For more information on this dataset, we refer to https://doi.org/10.25822/xmwv-yz91 and https://clinicaltrials.gov/ct2/show/NCT00642486.

MESA The *Multi-Ethnic Study of Atherosclerosis* (MESA) was a multi-ethnic longitudinal study of factors associated with the progression of cardiovascular disease across a cohort of black, white, Hispanic, and Chinese-American men and women aged 45-84 at study onset in 2000–2002 (X. Chen et al. 2015; G.-Q. Zhang et al. 2018). Between 2010–2012, 2237 participants further enrolled in the *MESA Sleep* sub-study and underwent (among others) overnight unattended PSG. We had 2056 subjects available for our study. EEG and EOG signals were recorded at 256 Hz and hardware low-pass filtered at 100 Hz. Hypnograms were scored according to the AASM criteria. For more information, we refer to https://doi.org/10.25822/n7hq-c406.

MROS A sub-study of the larger *Osteoporotic Fractures in Men* (MrOS) study investigated the association between sleep patterns, sleep-disordered breathing and cognition in community-dwelling men aged 67 and above who were not selected on the basis of sleep disorders or cognitive impairment (Blackwell et al. 2011; Song et al. 2015; G.-Q. Zhang et al. 2018). Subjects were enrolled from 6 different US clinical sites in Alabama, Minnesota, Pennsylvania, Oregon, and California. Between 2003-2005, 3135 subjects enrolled of which 2909 underwent in-home overnight polysomnography (PSG). In this study, we considered a total of 3926 PSG records (2900 from visit 1 and 1026 from visit 2) from 2903 subjects. We excluded a total of 7 records (IDs `aa2180`, `aa3370`, `aa1367`, `aa1715`, `aa1900`, `aa3903`, `aa3411` all from visit 1) due to missing EOG channels and/or sleep stage annotation files. EEG and EOG signals were recorded at 256 Hz and hardware high-pass filtered at 0.15 Hz. Hypnograms were scored according to the AASM criteria. For more information on the MROS dataset and studies, please refer to `https://doi.org/10.25822/kc27-0425`.

PHYS The over-night PSG data the from *2018 PhysioNet/CinC Challenge* were contributed by the Massachusetts General Hospital's Computational Clinical Neurophysiology Laboratory and the Clinical Data Animation Laboratory. The full dataset spans 1,985 patients who were monitored for the diagnosis of sleep disorders. The original challenge was automatic detection of arousal, but sleep stages were annotated by clinical staff. The dataset was split into two equal sized halves for training and testing. In our study we considered the 994 subjects publicly available in the training subset. EEG and EOG signals were recorded at 200 Hz. Hypnograms were scored according to the AASM criteria by a total of 7 annotators (1 scoring per PSG). For more information, we refer to `https://physionet.org/content/challenge-2018` and (Ghassemi et al. 2018; Goldberger et al. 2000).

SEDF-SC & SEDF-ST The *Sleep-EDF Database (Expanded)* consists of 197 whole-night PSG recordings. In the Sleep Cassette (SEDF-SC) sub-study, 153 PSGs were collected between 1987–1991 from healthy Caucasians aged 25–101 not taking sleep-related medication. The Sleep Telemetry (SEDF-ST) sub-study investigated the effect of temazepam intake on sleep in 22 Caucassian males and females. Participants took no other medication and were generally healthy but having mild difficulties falling asleep. Two recordings were collected from each individual on two nights at the hospital, one after temazepam intake and the other one after placebo intake. EEG and EOG signals were recorded at 100 Hz. Hypnograms were scored according to the Rechtschaffen and Kales criteria, which we aligned to AASM as described in the Methods section. The SEDF-SC database has been regularly used for benchmarking of automatic sleep stage classification algorithms. For more information on either sub-study, we refer to `https://doi.org/10.13026/C2C30J` and Goldberger

et al. (2000) and B. Kemp et al. (2000).

SHHS The *Sleep Heart Health Study* (SHHS) was a large, prospective cohort study investigating sleep-disordered breathing such as OSA as risk-factors for the development of cardiovascular disease (Quan et al. 1998; G.-Q. Zhang et al. 2018). A total of 6441 subjects were recruited from 6 already on-going National Heart, Lung, and Blood Institute studies (see `https://clinicaltrials.gov/ct2/show/NCT00005275` for details). Adults of age 40 or older, who were able and willing to undergo home PSG, were enrolled between 1995–1998. Between 2001-2003, a second PSG was obtained for 3295 of the participants. For our study, we had a total of 8444 PSG records available (5793 visit 1; 2651 visit 2) collected from 5797 individuals. EEG and EOG signals were recorded at 125 Hz and 50 Hz, respectively, and hardware high-pass filtered at 0.15 Hz. Hypnograms were scored according to the Rechtschaffen and Kales criteria, which we aligned to AASM as described in the Methods section. For more information on SHHS, we refer to `https://doi.org/10.25822/ghy8-ks59`.

SOF A sub-study of the larger *Study of Osteoporotic Fractures* (SOF) investigated the association between sleep-disordered breathing and cognitive impairment in community-dwelling Caucasian women aged 65 and above (Cummings et al. 1990; Spira et al. 2008; G.-Q. Zhang et al. 2018). Subjects were enrolled from four US cities between 1986–1988. An additional cohort of African-American women were recruited between 1997–1998. In our study, we considered the unattended, whole-night, in-home PSG data collected between 2002–2004 from 461 participants at SOF visit 8 (subjects originally enrolled from US metropolitan areas Minneapolis, Minnesota and and Pittsburgh, Pennsylvania between 1986-1988). EEG and EOG signals were recorded at 128 Hz and hardware high-pass filtered at 0.15 Hz. Hypnograms were scored according to the Rechtschaffen and Kales criteria, which we aligned to AASM as described in the Methods section. For more information, we refer to `https://doi.org/10.13026/C2X676`.

DCSM This new dataset was collected and prepared by the Danish Centre for Sleep Medicine (DCSM) and comprises 255 whole-night PSG recordings of patients visiting the center for diagnosis of non-specific sleep related disorders. The records are fully anonymized and were selected randomly. The included subjects thus likely vary in demographic characteristics, diagnostic background and sleep/non-sleep related medication usage. The PSGs were collected between 2015-2018. EEG and EOG signals were recorded at 256 Hz and bandpass filtered to interval 0.3 Hz - 70 Hz (3dB limits). Hypnograms were scored according to the AASM criteria. This dataset serves as an unbiased, random sample from the distribution of data generated by the DCSM. The DCSM dataset is publicaly available at `https://sid.erda.dk/wsgi-bin/ls.py?share_id=fUH3xbOXv8`. This repository will be frozen and issued a DOI for persistent access following the review process.

ISRUC-SG1, ISRUC-SG2 & ISRUC-SG3 The ISRUC dataset consists of randomly selected all-night PSG recordings acquired by the Sleep Medicine Centre of the Hospital of Coimbra University, Portugal (Khalighi et al. 2016). It covers both healthy subjects and patients with sleep disorders under the effect of sleep medication. It is divided into three sub-groups (ISRUC-SG1, -SG2, -SG3) with 100 sleep disordered adults, 8 sleep disordered adults with PSGs aqquired twice on different nights, and 10 healthy control subjects in each of the three sub-groups, respectively. Data were aqquired between 2009–2013. All records were scored by two experts. We considered hypnograms from `annotator 1` for all records but `subject_2_visit_2` of `ISURC-SG2` for which we used the hypnogram of `annotator 2` due to missing data. EEG and EOG signals were recorded at 200 Hz and filtered using a bandpass Butterworth filter with lower and higher cutoff frequencies of 0.3 Hz and 35 Hz, respectively. Hypnograms were scored according to the AASM criteria. For more information on the ISRUC dataset, we refer to `https://sleeptight.isr.uc.pt`.

MASS-C1 & MASS-C3 The *Montreal Archive of Sleep Studies* (MASS) pooled 200 whole-night recordings from three different hospital-based sleep laboratories of the Center for Advanced Research in Sleep Medicine, Montreal, Canada (O'Reilly et al. 2014). Subjects were between 18–76 years at the time of recording, which occurred in the period 2001–2013. The subjects were organized into five subsets (C1-C5) according to the research protocols used for data acquisition. All included subjects were healthy controls, although an apnea-hypnea index of up to 20 (moderate sleep apnea) was allowed for subjects in C1. In this study, we considered PSG recordings from subsets C1 (53 subjects) and C3 (62 subjects) for which the experts annotated 30-second intervals in line with the other datasets. EEG and EOG signals were recorded at 256 Hz and hardware low-passed filtered at 0.10 Hz (EOG) or 0.30 Hz (EEG) and high-pass filtered at 30 Hz (EOG) or 100 Hz (EEG). Hypnograms were scored according to the AASM criteria. For more information, we refer to `http://ceams-carsm.ca/en/MASS`.

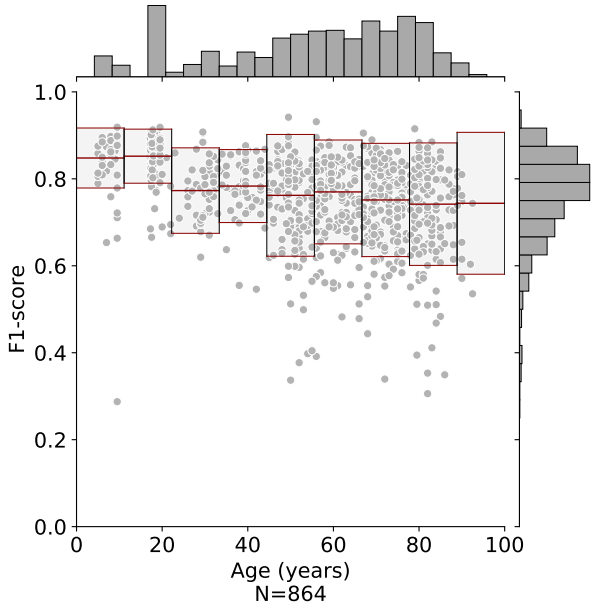
SVUH The *St. Vincent's University Hospital / University College Dublin Sleep Apnea Database* (SVUH) contains 25 full overnight PSG records of randomly selected individuals under diagnosis for either obstructive sleep apnea, central sleep apnea or primary snoring (Goldberger et al. 2000). Subjects were enrolled over a 6-month period. We considered data from the revised database of 2001. Subjects were at least 18 years old and had no known cardiac disease, had no autonomic dysfunction, and took no medication known to interfere with heart rate. EEG and EOG signals were recorded at 128 Hz. Hypnograms were scored according to the Rechtschaffen and Kales criteria, which we aligned to AASM as described in the Methods section. For more information, we refer to `https://doi.org/10.13026/C26C7D`.

DOD-H & DOD-O The *Dreem Open Dataset – Healthy* (`DOD-H`) was collected from 25 volunteers at the French Armed Forces Biomedical Research Institute's Fatigue and Vigilance Unit in France. Subjects were without sleep complaints, aged 18-65 and locally recruited without regard to gender or ethnicity. The *Dreem Open Dataset – Obstructive* (`DOD-O`) was collected from the Stanford Sleep Medicine Center, California, US from 55 patients (clinical trial NCT03657329) with clinical suspicion for sleep-related breathing disorder. Individuals clinically diagnosed with sleep disorders other than OSA, suffering from morbid obesity, taking sleep medications or with certain cardiopulmonary or neurological comorbidities were excluded from the study. EEG and EOG signals from both `DOD-H` and `DOD-O` were sampled at 256 Hz and each PSG was scored by 5 individual experts from 3 different sleep clinics. All experts were registered Sleep Technologists with at least 5 years of clinical scoring experience. For more information on the `DOD` datasets and consensus scoring, we refer to the publications of Arnal et al. (2019), Guillot et al. (2019), and Thorey et al. (2019).

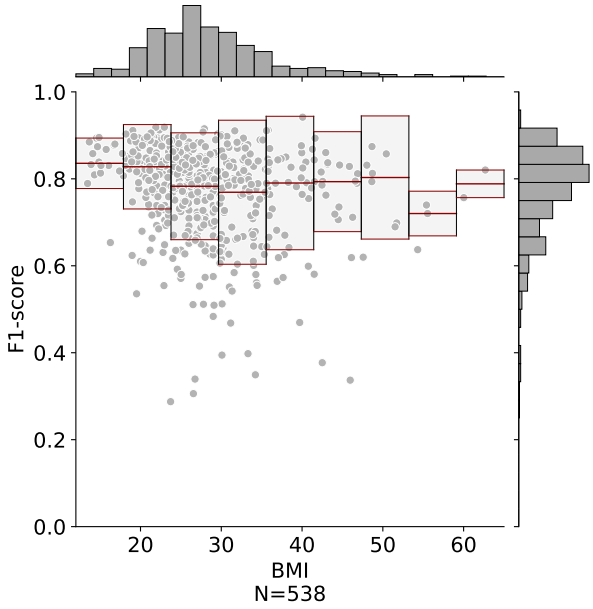
D.2 Supplementary Note: Demographic Bias

We conducted an analysis of potential demographic bias in the average U-Sleep performance. We considered the variables age, sex and BMI and accounted for dataset origin. We could not evaluate important variables such as disease state and ethnicity, because the required information was missing was missing for several datasets. Supplementary Figure D.1 shows graphical representations of the test-set distribution of F1 scores as a function of age, sex, BMI and general disease stage, respectively. Note that these plots show only correlations, not causal relations.

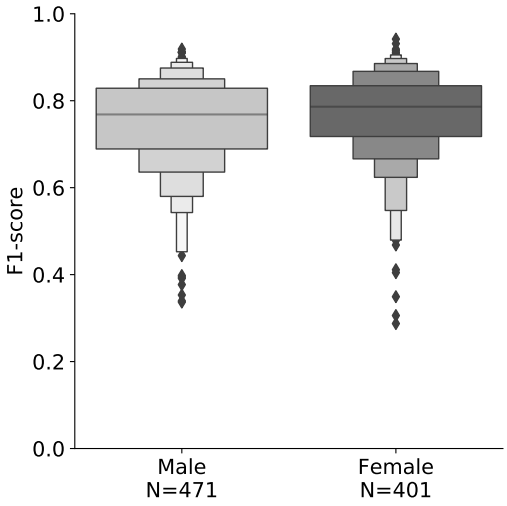
We fitted a beta regression model (using the `betareg`, Cribari-Neto et al. (2010), v3.1-3 package in `R`, Team (2019), v3.6.1) on 532 records from the test sets of datasets ABC, CCSHS, CFS, CHAT, HPAP, MROS, SHHS, SOF and SVUH. The 532 records represent all available test-set records for which we have age, sex and BMI information available. The regression model predicts the mean F1 score as a function of those parameters along with variables encoding the dataset origin of each sample giving a total 11 covariates. The estimated coefficients of the model were -0.004 ± 0.007 for BMI (95% CI, $z = -1.273$, $p = 0.203$), -0.012 ± 0.004 for age (95% CI, $z = -6.141$, $p < 0.001$), and -0.102 ± 0.102 for sex (difference if subject is Male, 95% CI, $z = -1.954$, $p = 0.051$). Coefficients testing were done using two-sided Z-tests. The interpretation of the coefficients is that the expected F1 performance drops with increasing BMI and increasing age as well as for male subjects. However, only age was significant ($p < 0.05$). It is likely that this observation is confounded by the general worsening of health with age.



(a) Majority vote mean F1-scores by subject age.

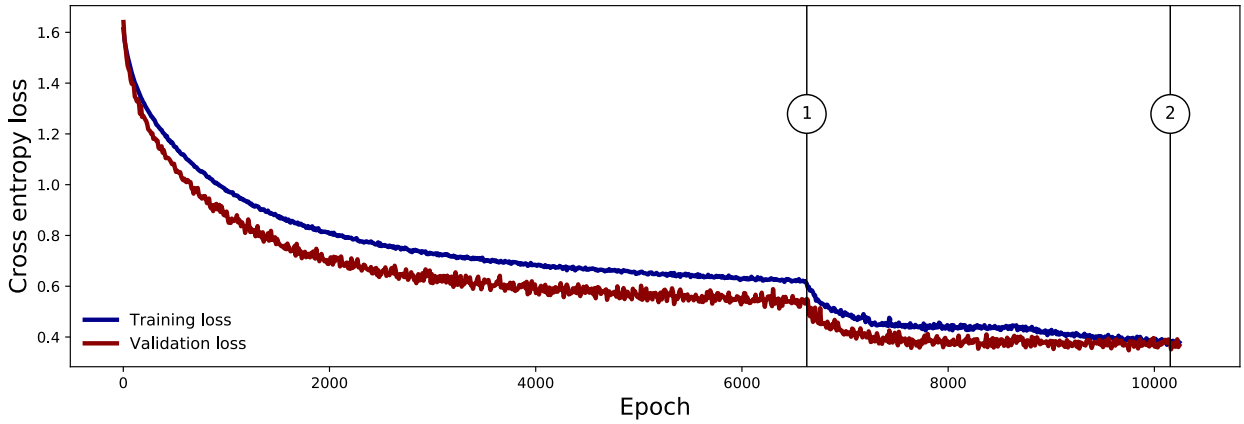


(b) Majority vote mean F1-scores by subject BMI.

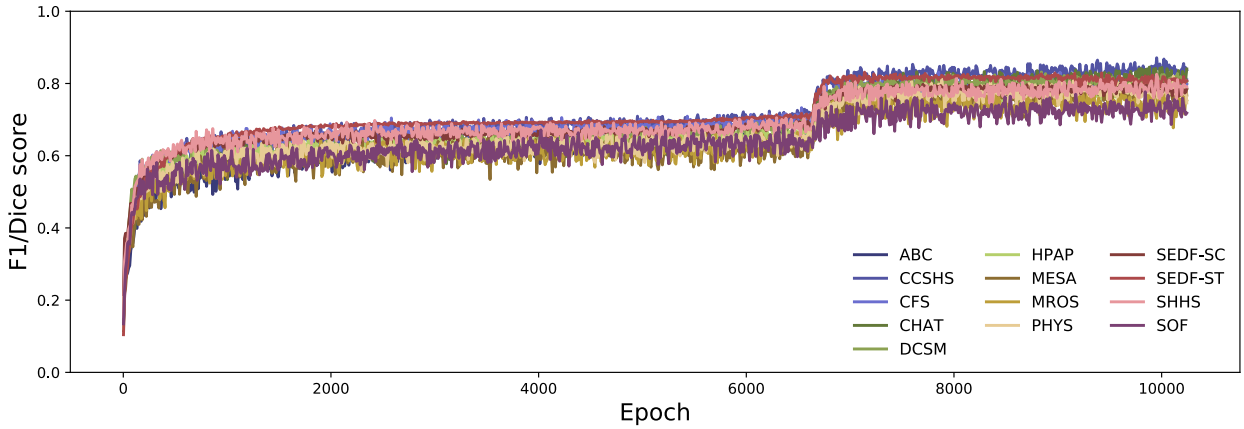


(c) Majority vote mean F1-scores by subject sex.

Figure D.1: Correlations between U-Sleep per-subject mean F1 score performance (using majority votes) and individual demographic variables. In panels (a) and (b) the red center lines show median values in 10 equally sized bins. The lower and upper red lines together represent the interquartile range. In panel (c) letter-value plots (H. Hofmann et al. 2017) visualize the median (center black line) and 10 other quantiles (letter-values, specifically). Observations beyond the most extreme letter-values are labeled outliers and plotted as diamond shapes. Note that the widths of each box in the letter-value plots are arbitrary and serve only to visually separate individual boxes. Please refer to the Supplementary Note: Demographic Bias for statistical analysis and discussion.



(a) The training and validation loss on the training and validation set. Little to no overfitting (reduced training loss with stagnant or increasing validation loss) is observed. Two points of interest are marked: 1) The sudden improvement in performance occurs as the model starts improving on the difficult N1 sleep stage. Up until this point, U-Sleep would rarely predict N1 stages at all, resulting in a lower mean performance. 2) The finally selected model at epoch 10154. Training for 150 epochs after this point did not further improve validation performance.



(b) Mean F1 score computed across all datasets using random subsets of the validation data after each epoch of training. The mean F1 increases steadily over time on all datasets, indicating that the model is able to simultaneously learn the function across all clinical cohorts.

Figure D.2: U-Sleep learning curves. It took a total of $\approx 4,500,000$ gradient updates (processed batches of data) to train the model to convergence, equivalent to observing $\approx 9,582$ years of (non-unique) annotated PSG data. The total training set length is ≈ 19.4 years. The long training time needed to obtain the final model is a result of both the highly challenging task – learning sleep staging across clinical cohorts with varying and noisy labels, randomly varying input channels as well as augmentation – and that we chose to train U-Sleep using a very small learning rate (please refer to the Methods section). As we were interested only in a single, final version of the U-Sleep model, the long training time is only an issue because of the energy consumption. We estimate that training U-Sleep consumed up to a total of $1,121$ kWh (96.1 kg CO_2 eq.) using the `CarbonTracker` tool (Anthony et al. 2020) with an added 25 % margin.

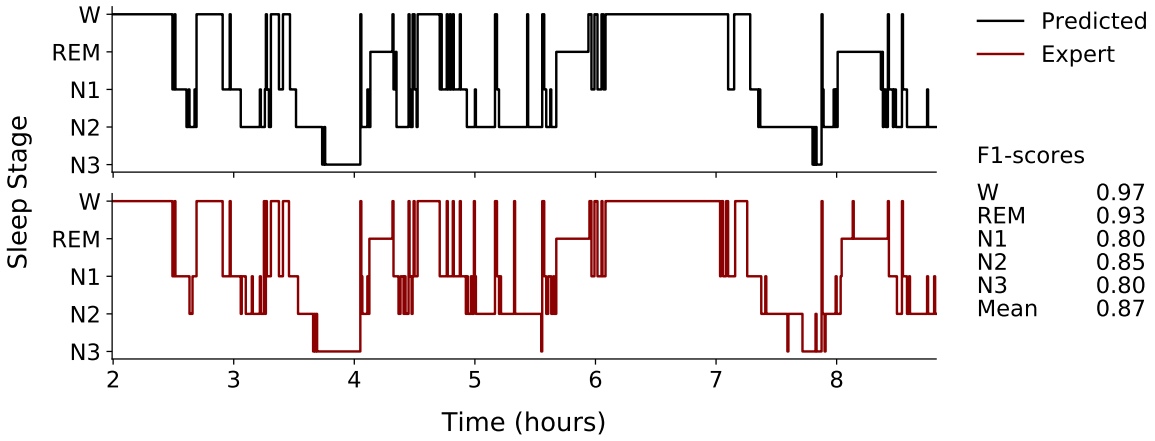
Table D.1: U-Sleep model topology for input window size $i = 3840$ (30 seconds of 128 Hz signal), number of input channels $C = 2$, sequence length $T = 35$, number of output classes $K = 5$ and complexity factor scaling value $\alpha = 1.67$. The complexity scaling modifies the filter number in each block or layer as described in the Methods section to a number $c' = \lfloor c \cdot \sqrt{\alpha} \rfloor$ where c is the original number of filters. Note that $T \cdot i = 134400$. Each encoder block performs the following operations: convolution → activation function → batch normalization → (zero-padding if input length is odd) → max pooling (kernel width 2, stride 2). Each encoder block also outputs a residual connection (the output of the layer immediately before max pooling) which is passed to its corresponding decoder block. Each decoder block performs the following operations on its input: nearest-neighbour up-sampling (kernel width 2) → convolution → activation function → batch normalization → (crop if needed to match residual connection input) → concatenation with residual connection → convolution → activation function → batch normalization. The average pooling layer (ID=28) has striding of $i = 3840$.

ID	Layer Type	Output dim	Kernel	Filters	Activation	Pad
-	Input (symbolic)	$T \cdot i \times C$	-	-	-	-
1	Input	134400×2	-	-	-	-
2	Encoder Block	67200×6	9	6	ELU	same
3	Encoder Block	33600×9	9	9	ELU	same
4	Encoder Block	16800×11	9	11	ELU	same
5	Encoder Block	8400×15	9	15	ELU	same
6	Encoder Block	4200×20	9	20	ELU	same
7	Encoder Block	2100×28	9	28	ELU	same
8	Encoder Block	1050×40	9	40	ELU	same
9	Encoder Block	525×55	9	55	ELU	same
10	Encoder Block	263×77	9	77	ELU	same
11	Encoder Block	132×108	9	108	ELU	same
12	Encoder Block	66×152	9	152	ELU	same
13	Encoder Block	33×214	9	214	ELU	same
14	Convolution + Batch Norm.	33×302	9	306	ELU	same
15	Decoder Block (res=13)	66×428	9	214	ELU	same
16	Decoder Block (res=12)	132×304	9	152	ELU	same
17	Decoder Block (res=11)	264×216	9	108	ELU	same
18	Decoder Block (res=10)	526×154	9	77	ELU	same
19	Decoder Block (res=9)	1050×110	9	55	ELU	same
20	Decoder Block (res=8)	2100×80	9	40	ELU	same
21	Decoder Block (res=7)	4200×56	9	28	ELU	same
22	Decoder Block (res=6)	8400×40	9	20	ELU	same
23	Decoder Block (res=5)	16800×30	9	15	ELU	same
24	Decoder Block (res=4)	33600×22	9	11	ELU	same
25	Decoder Block (res=3)	67200×18	9	9	ELU	same
26	Decoder Block (res=2)	134400×12	9	6	ELU	same
27	Convolution	134400×6	1	6	TanH	same
28	Average Pooling	35×6	3840	-	-	valid
29	Convolution	35×5	1	5	ELU	same
30	Convolution	35×5	1	5	Softmax	same

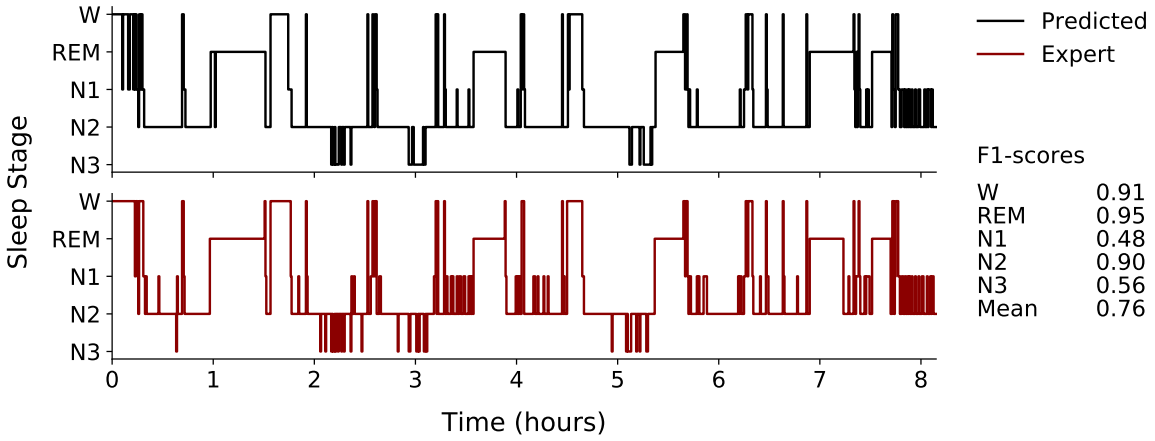
Trainable parameters: $3,114,337$

Table D.2: U-Sleep Model, Optimization and Pre-Processing Hyperparameters.

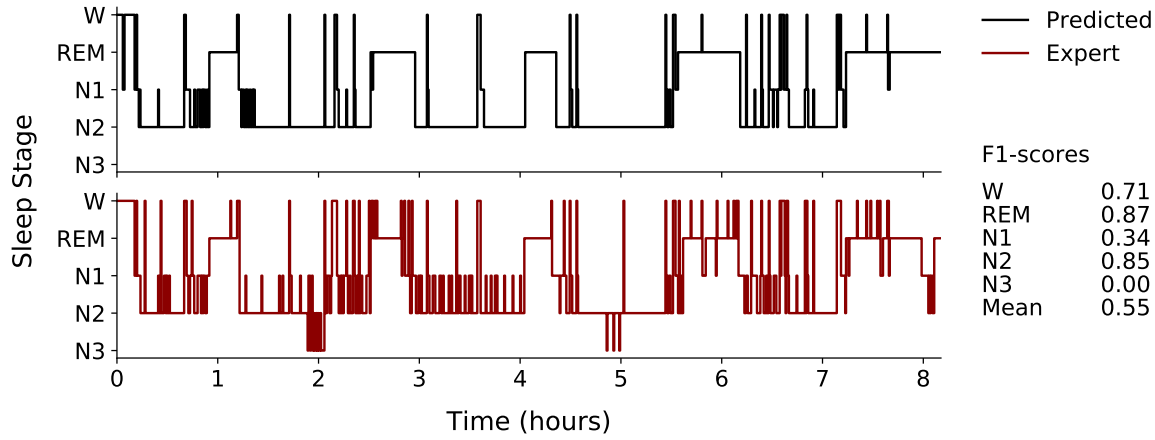
Parameter	Value	Notes
Optimizer	Adam	See Kingma et al. (2015).
Learning rate -	10^{-7}	
β_1 -	0.9	
β_2 -	0.999	
ϵ -	$1 \cdot 10^{-8}$	
Loss function	Cross-entropy	
Regularization -	None	
Class balancing -	None	
Base Topology	1D U-Net	The input dimensionality is the number of data points in a single PSG segment (one segment is 30 seconds in typical sleep staging, giving input dimensionality 3840 for sample rate $S = 128$). T is the number of contiguous segments the model operates on at once. T may be dynamically adjusted. Zero-padding is needed to decode to dimensions equal to the input (Ioffe et al. 2015; Odena et al. 2016; Ronneberger et al. 2015; Yu et al. 2016). ELU=Exponential Linear Units (Clevert et al. 2016).
Input dim. -	3840	
Window size (T) -	35	
Depth -	12	
Up-sampling -	Nearest neighbour	
Activations -	ELU	
Conv. kernel size -	9	
Conv. kernel dilation size -	9	
Max-pool kernel size -	2	
Padding -	True ('same')	
Batch normalization -	True	
Parameters -	$\approx 3.12 \cdot 10^6$	
Pre-processing	Robust scaling	Record- and channel-wise transformation to distribution of median 0 and IQR 1.
Post-processing	None	
Re-sampling (S)	128 Hz	Re-sampling uses polyphase filtering (implementation: `scipy.signal.resample_poly`, see Virtanen et al. 2020). Clamping of absolute values deviating from the median by more than 20 times the IQR of the channel.
Batch size (B)	64	For element in a batch, a class from the label set {W, N1, N2, N3, R} is determined by uniform sampling. A random PSG record of this class is sampled, from which the input window is sampled randomly so that the selected class is in the window.
Class sampling prob. -	Uniform	
Training epochs	∞	Training continues until 100 consecutive epochs without validation performance improvements. 443 steps amounts to roughly 10^6 30-second segments (or labels, equivalently).
Steps per epoch	443	
Early stopping criteria	Validation F1	Mean per-class F1 scores computed over random subsets of up to 20 validation records from each dataset.
Model selection criteria	Validation F1	



(a) Hypnogram with highest observed F1-score (record `abc-baseline-900026`).

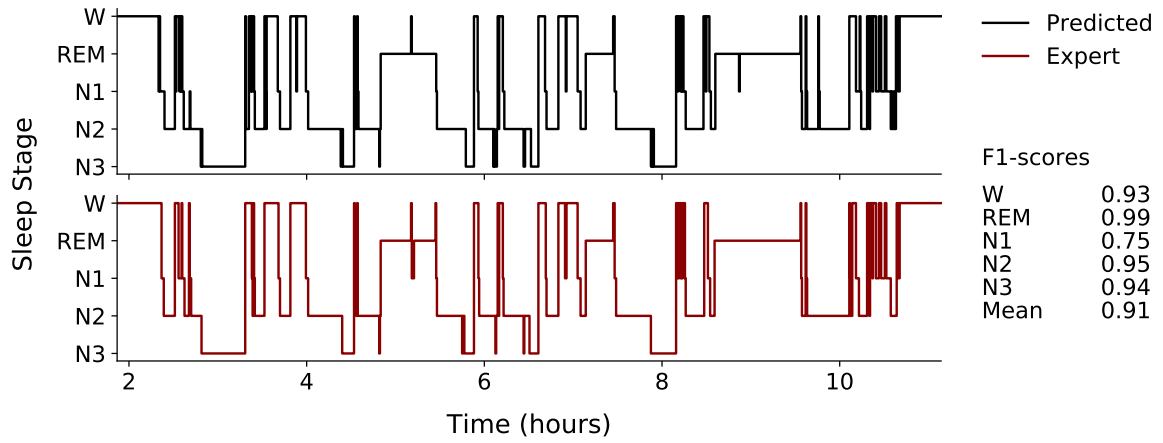


(b) Hypnogram with F1-score nearest dataset median (record `abc-baseline-900039`).

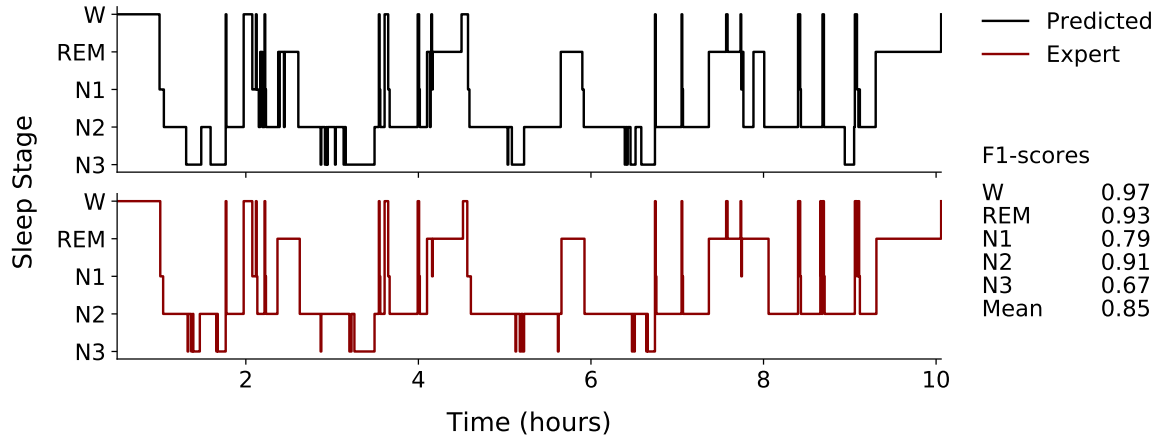


(c) Hypnogram with lowest observed F1-score (record `abc-baseline-900014`).

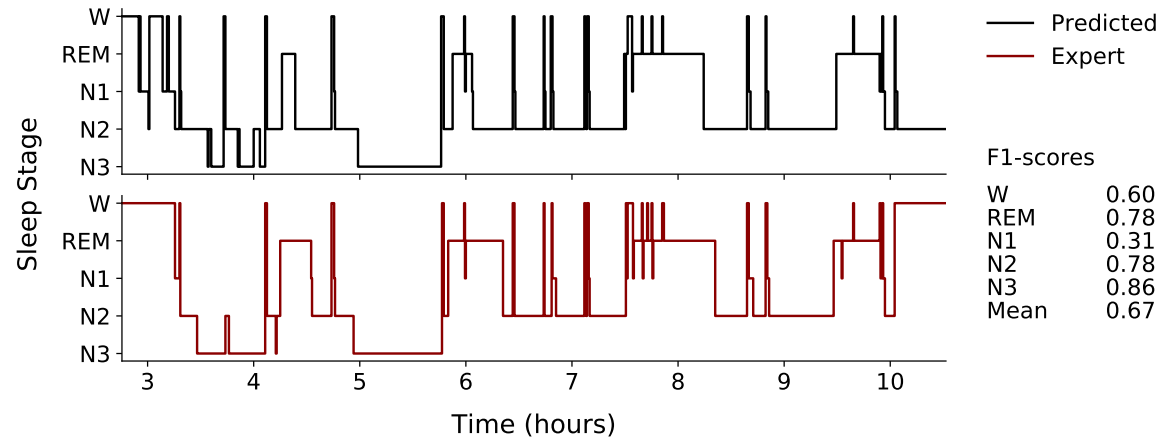
Figure D.3: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `ABC`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `ccshs-trec-1800544`).

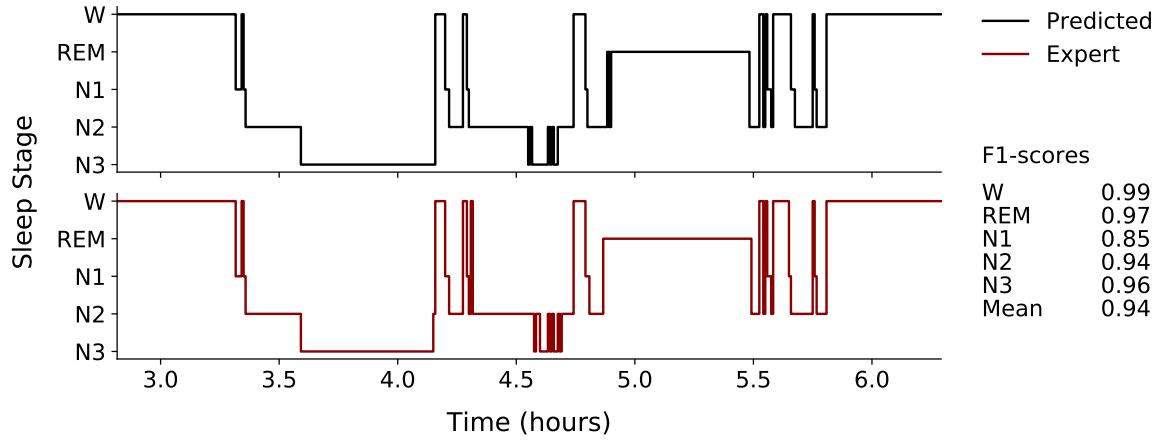


(b) Hypnogram with F1-score nearest dataset median (record `ccshs-trec-1800195`).

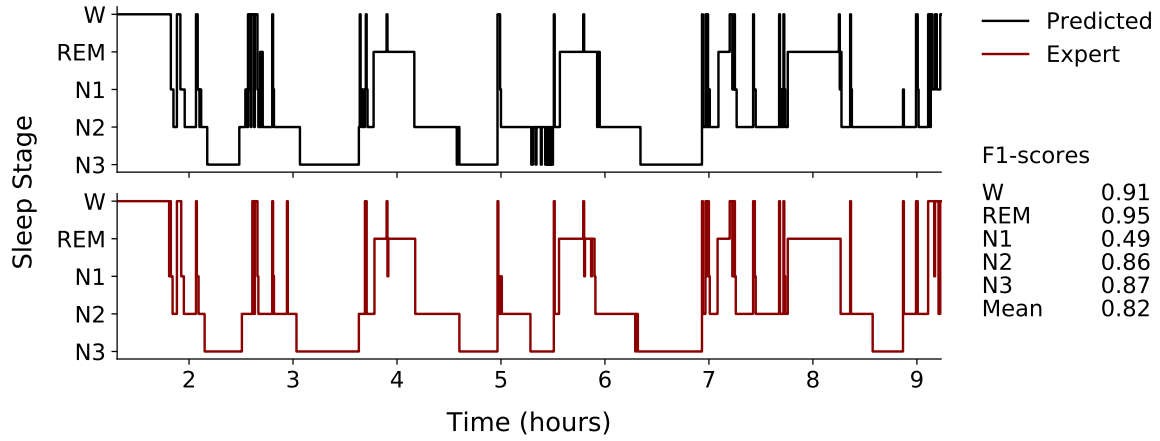


(c) Hypnogram with lowest observed F1-score (record `ccshs-trec-1800007`).

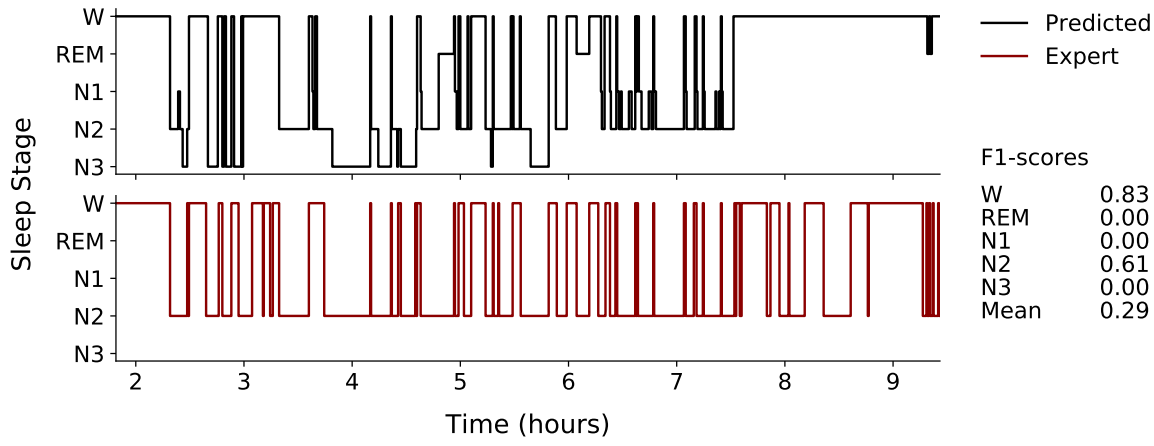
Figure D.4: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `CCSHS`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `cfs-visit5-802273`).

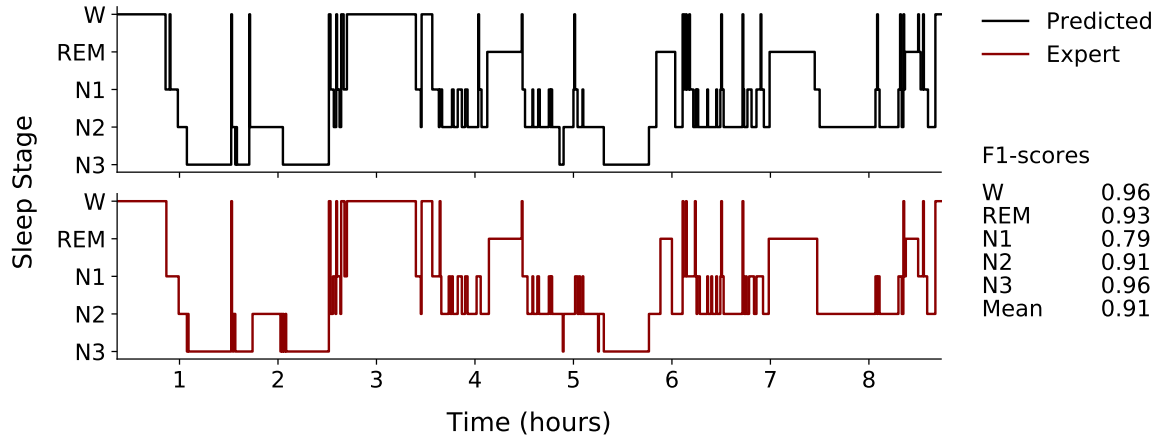


(b) Hypnogram with F1-score nearest dataset median (record `cfs-visit5-802658`).

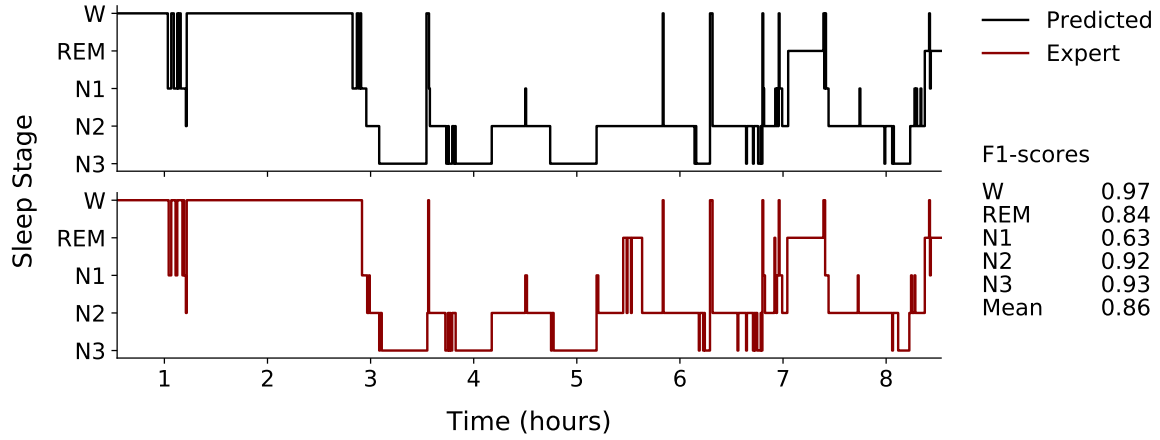


(c) Hypnogram with lowest observed F1-score (record `cfs-visit5-802299`).

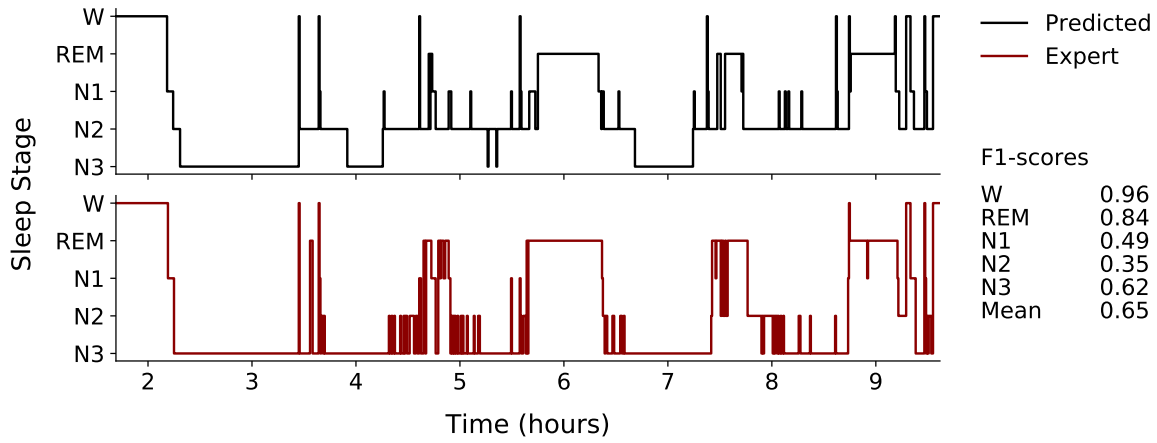
Figure D.5: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `CFS`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `chat-baseline-nonrandomized-300405`).

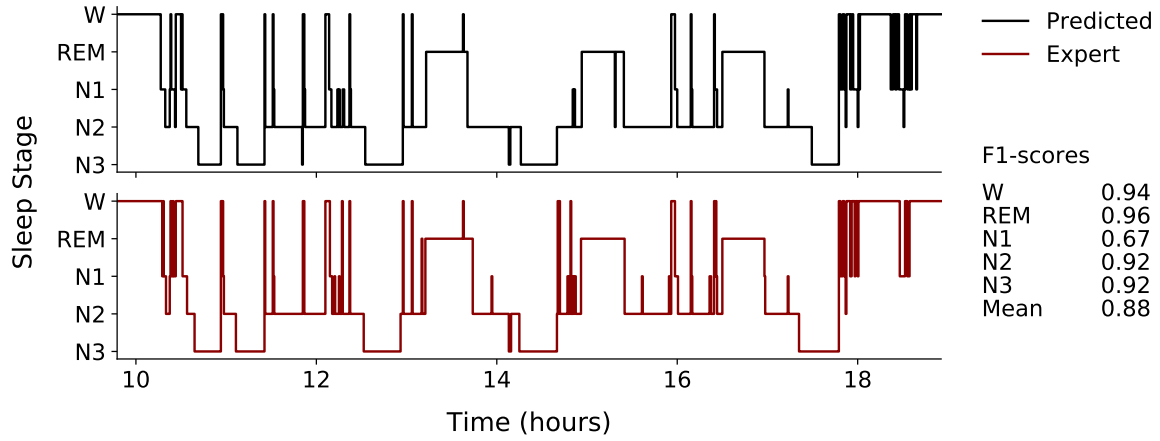


(b) Hypnogram with F1-score nearest dataset median (record `chat-baseline-nonrandomized-301034`).

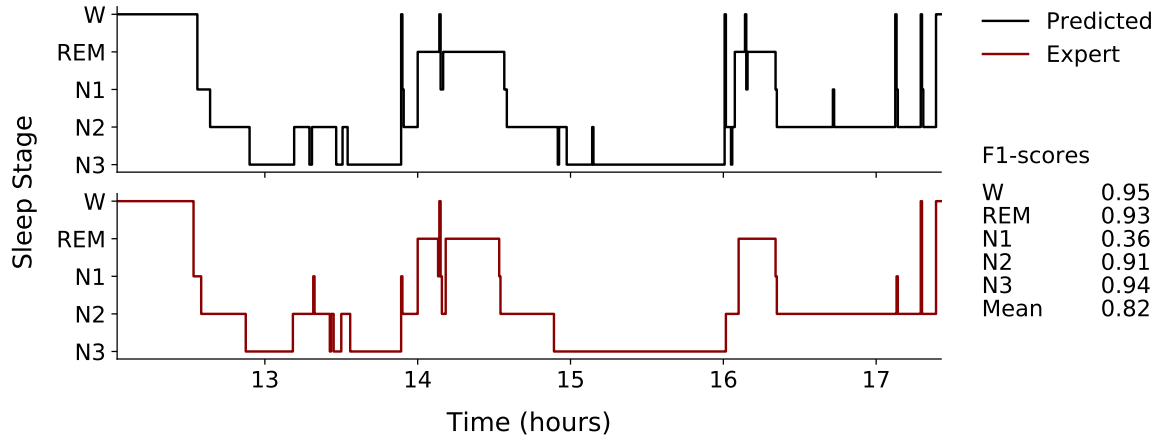


(c) Hypnogram with lowest observed F1-score (record `chat-baseline-300397`).

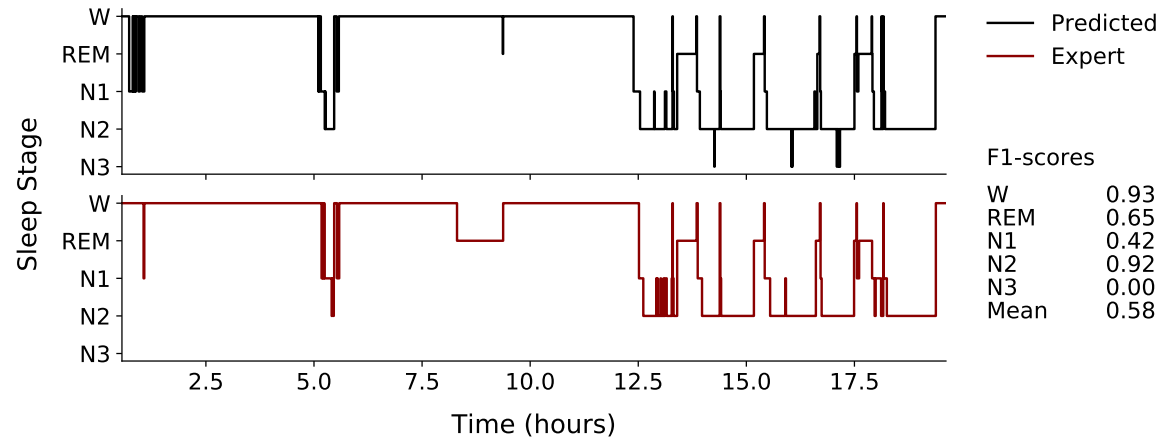
Figure D.6: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `CHAT`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `285ab4bdf51f`).

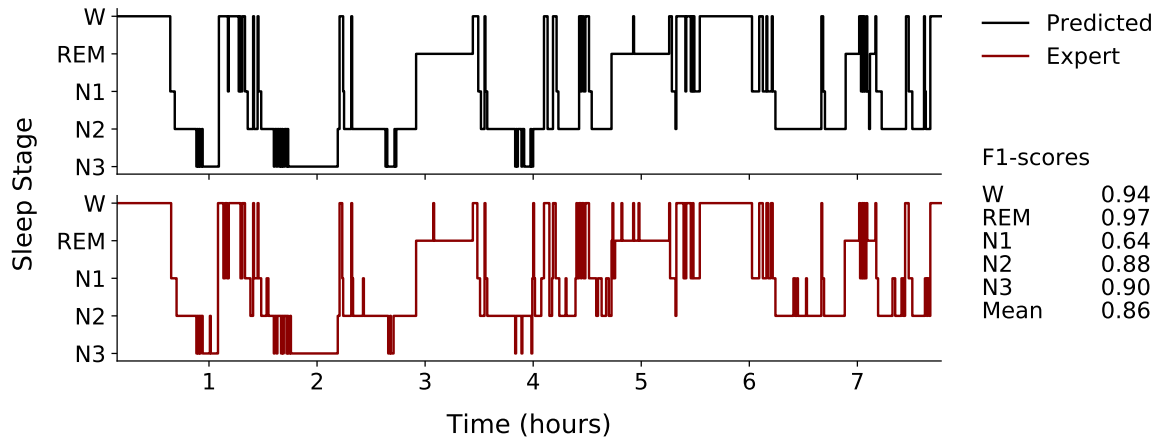


(b) Hypnogram with F1-score nearest dataset median (record `4dac221360bb`).

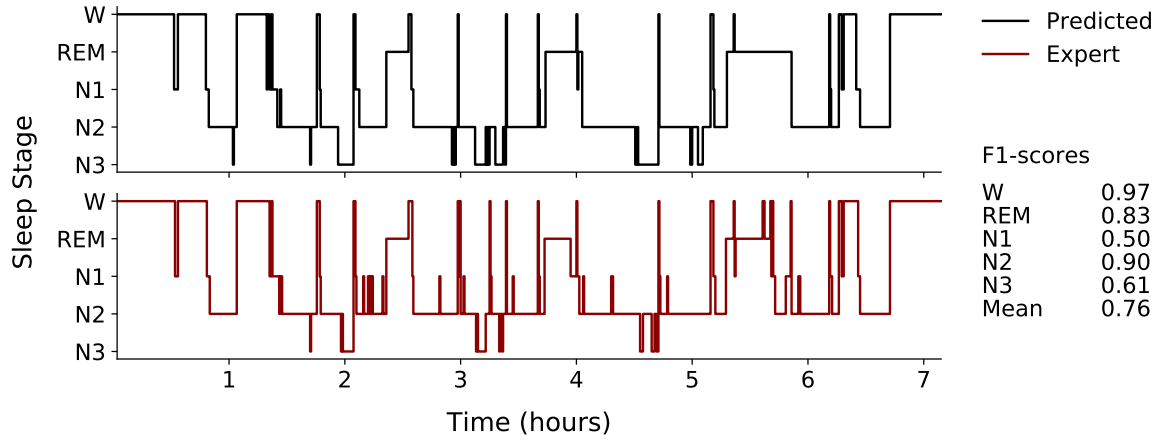


(c) Hypnogram with lowest observed F1-score (record `65fd36d709ae`).

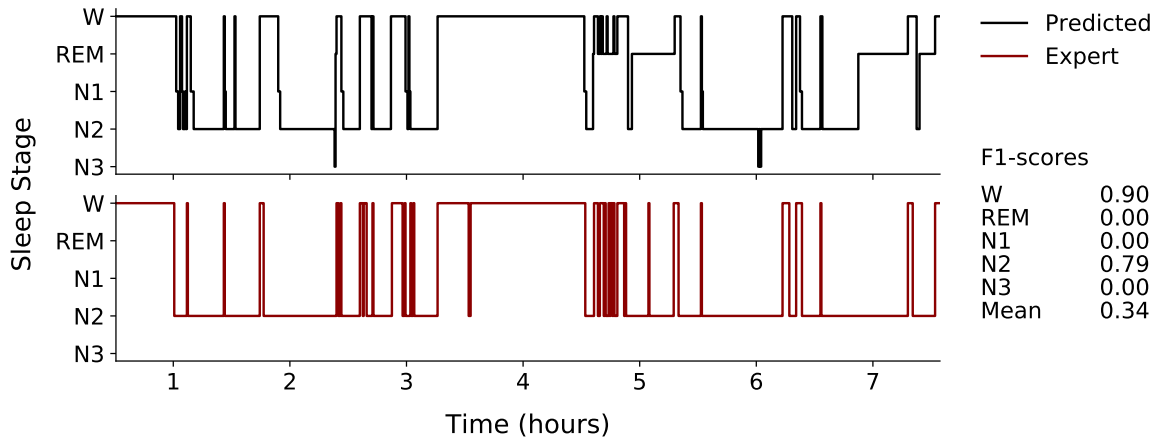
Figure D.7: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `DCSM`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `homepap-lab-full-1600255`).

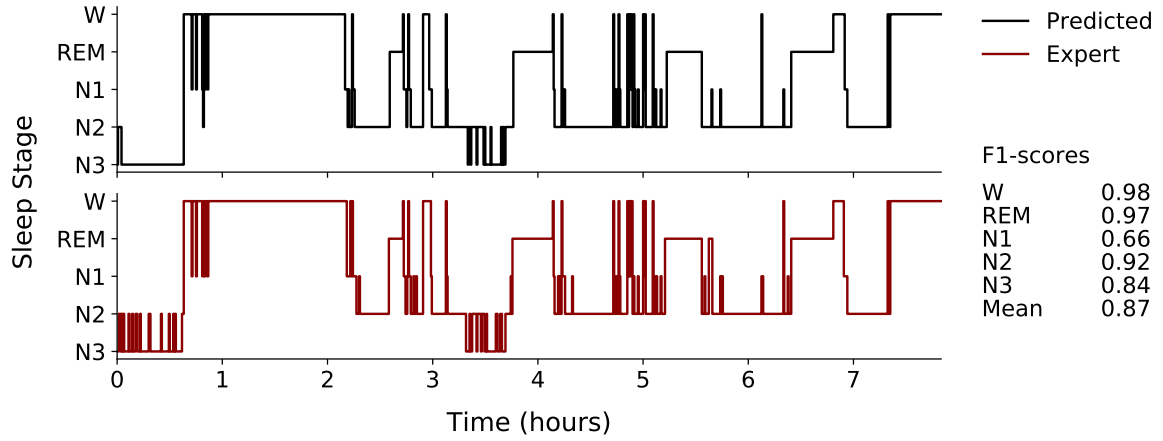


(b) Hypnogram with F1-score nearest dataset median (record `homepap-lab-full-1600319`).

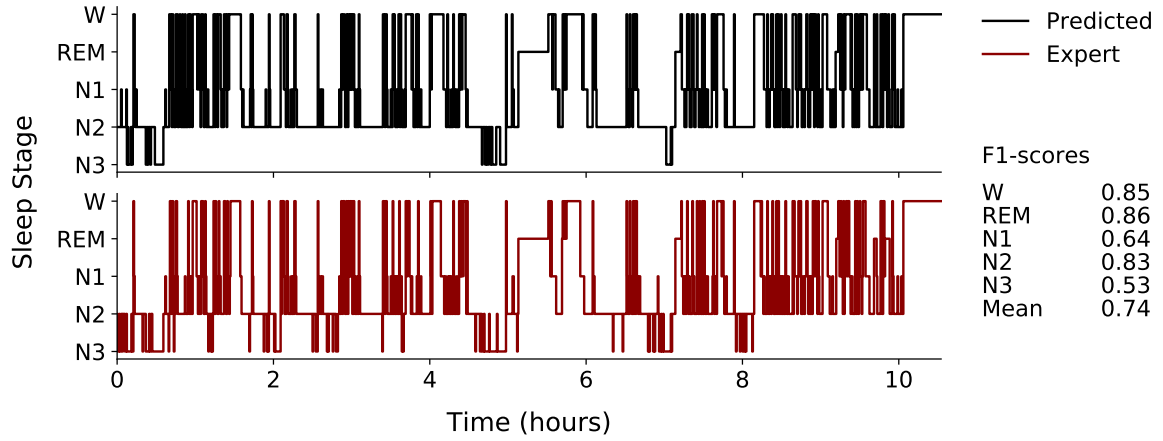


(c) Hypnogram with lowest observed F1-score (record `homepap-lab-split-1600251`).

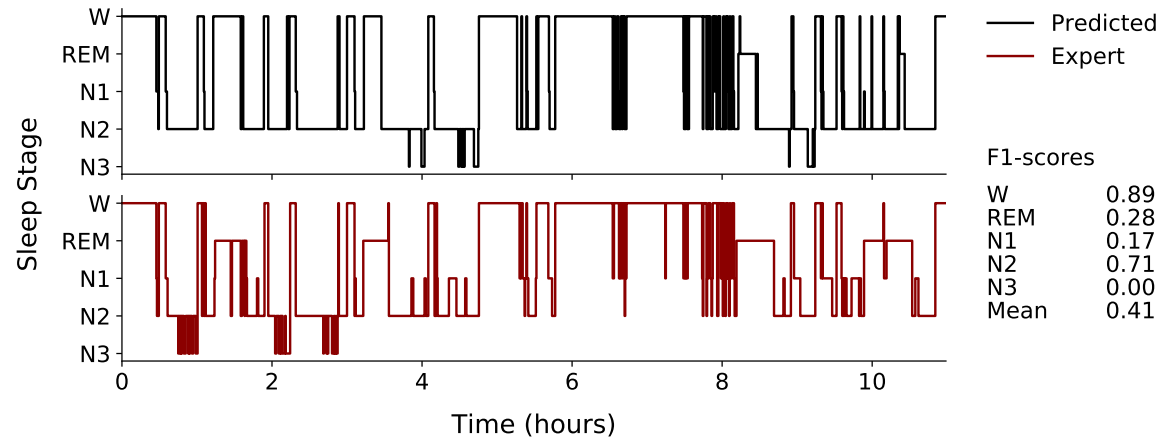
Figure D.8: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `HPAP`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `mesa-sleep-4682`).

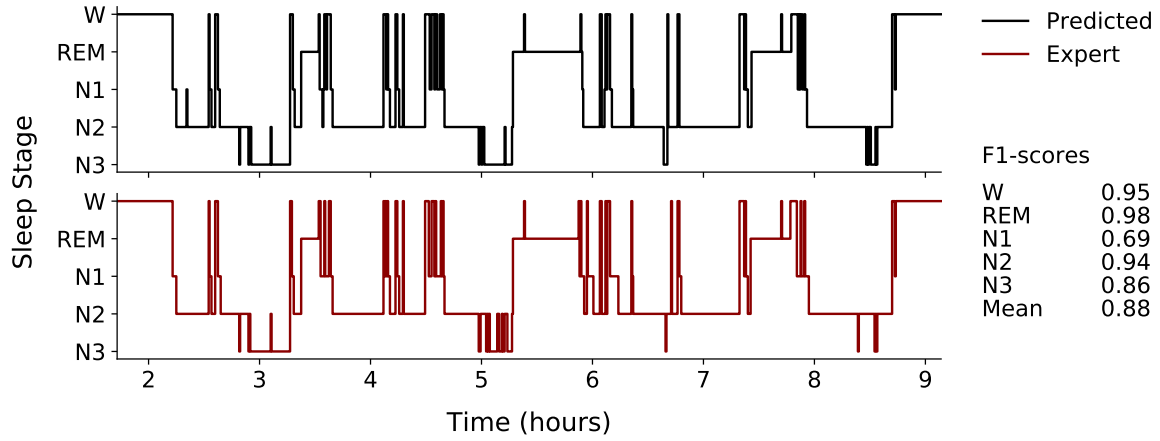


(b) Hypnogram with F1-score nearest dataset median (record `mesa-sleep-2834`).

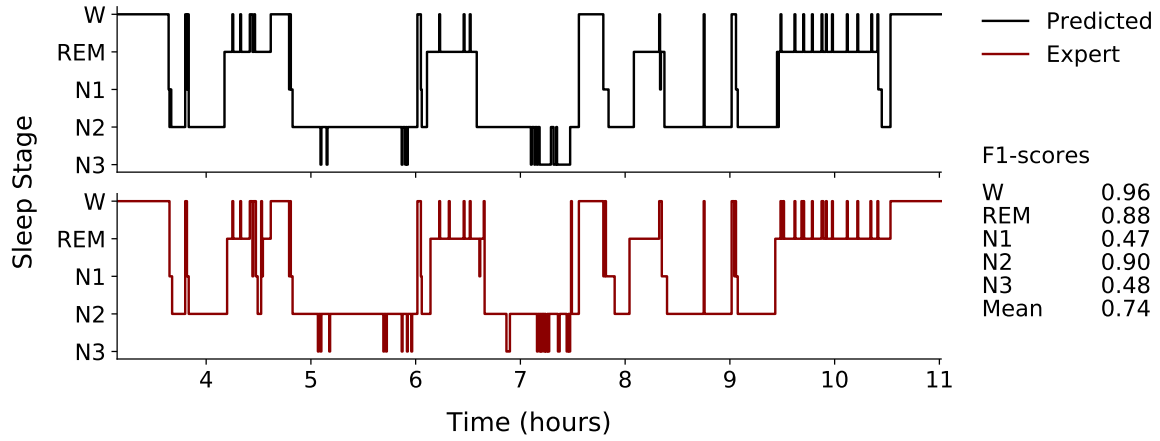


(c) Hypnogram with lowest observed F1-score (record `mesa-sleep-5680`).

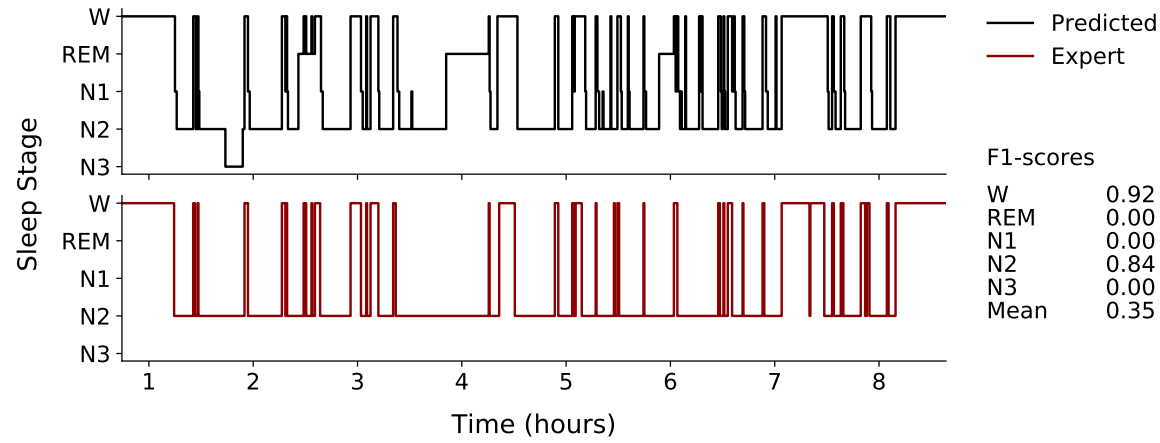
Figure D.9: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `MESA`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `mros-visit1-aa2023`).

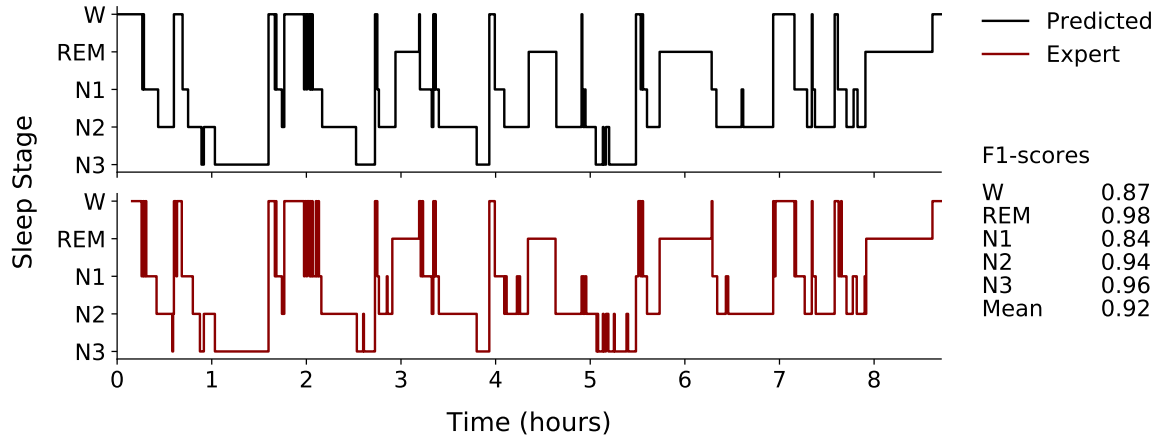


(b) Hypnogram with F1-score nearest dataset median (record `mros-visit1-aa2359`).

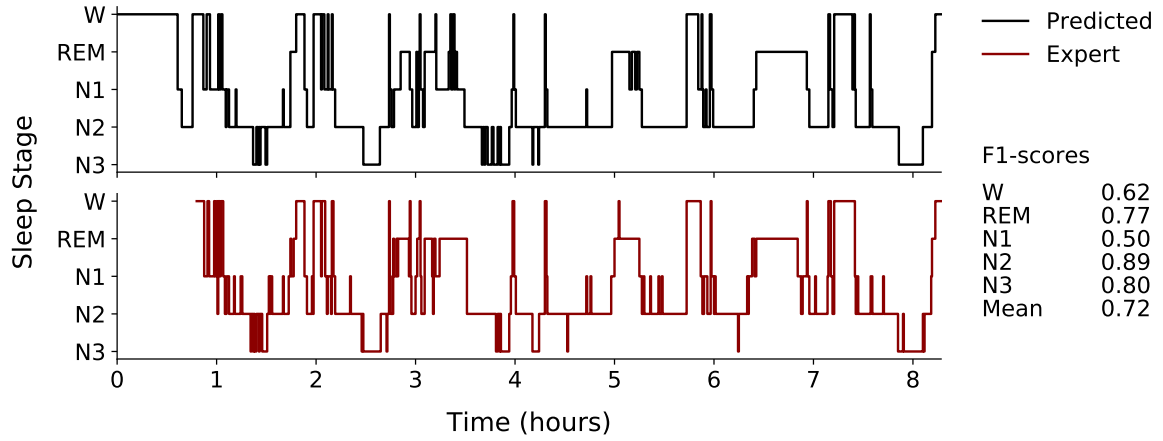


(c) Hypnogram with lowest observed F1-score (record `mros-visit2-aa3175`).

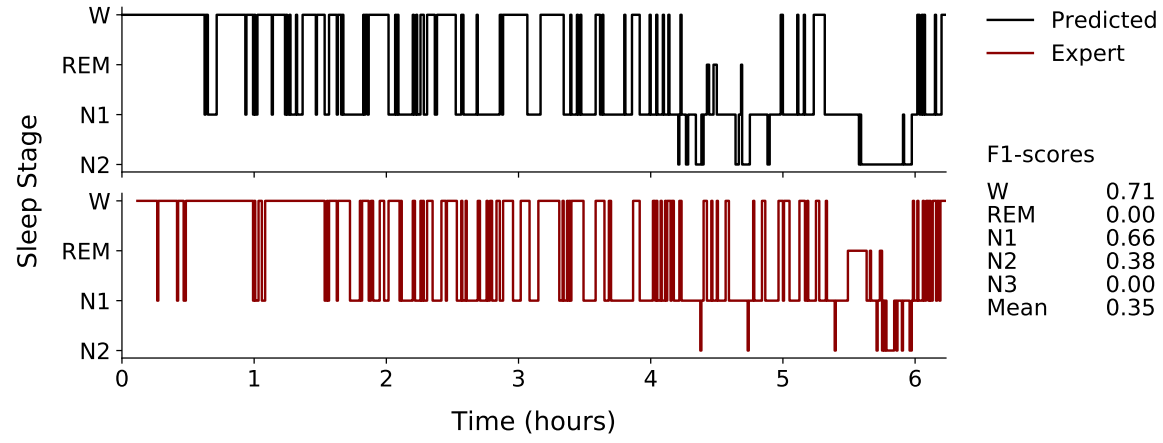
Figure D.10: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `MROS`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `tr07-0891`).

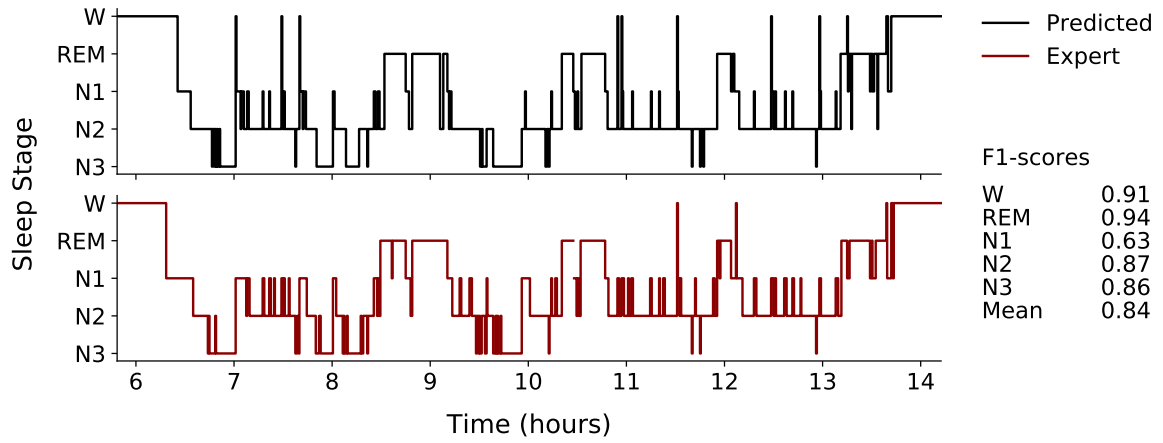


(b) Hypnogram with F1-score nearest dataset median (record `tr07-0394`).

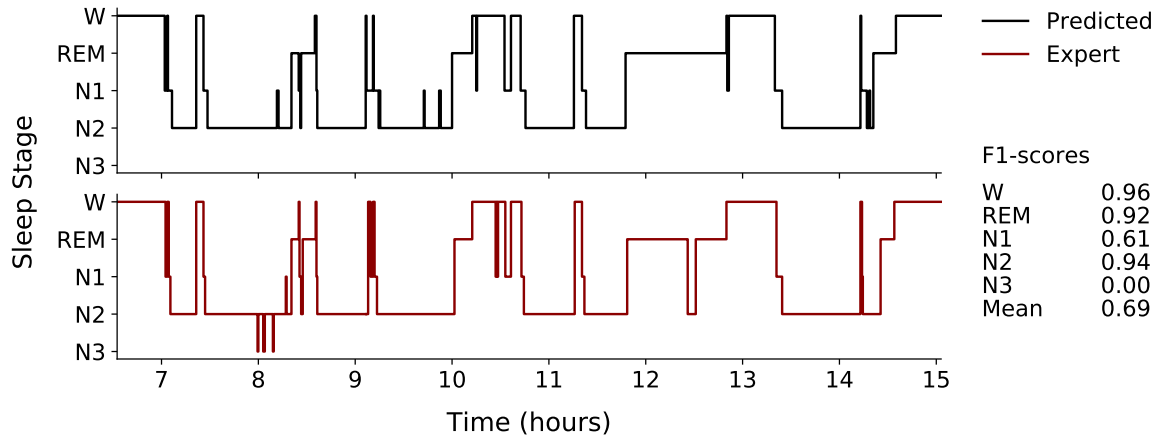


(c) Hypnogram with lowest observed F1-score (record `tr07-0828`).

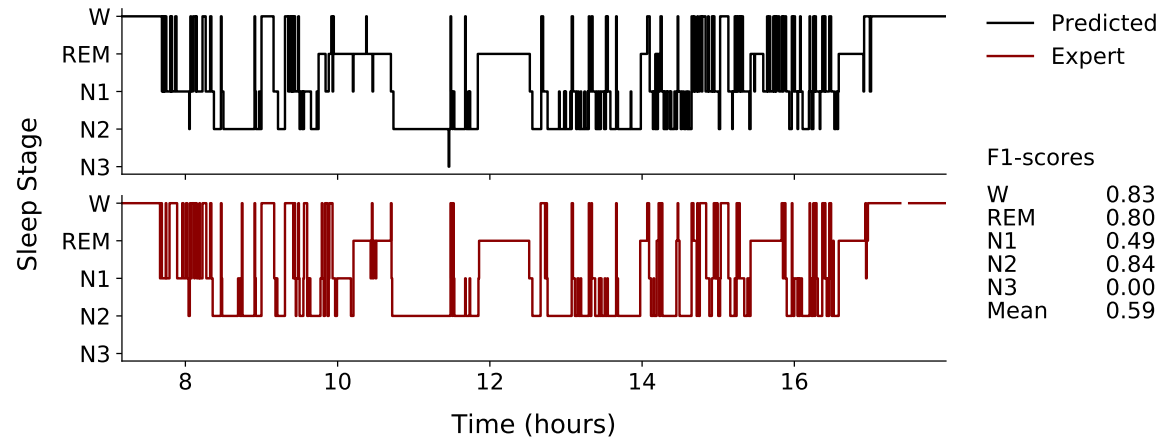
Figure D.11: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `PHYS`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `SC4022E0-PSG`).

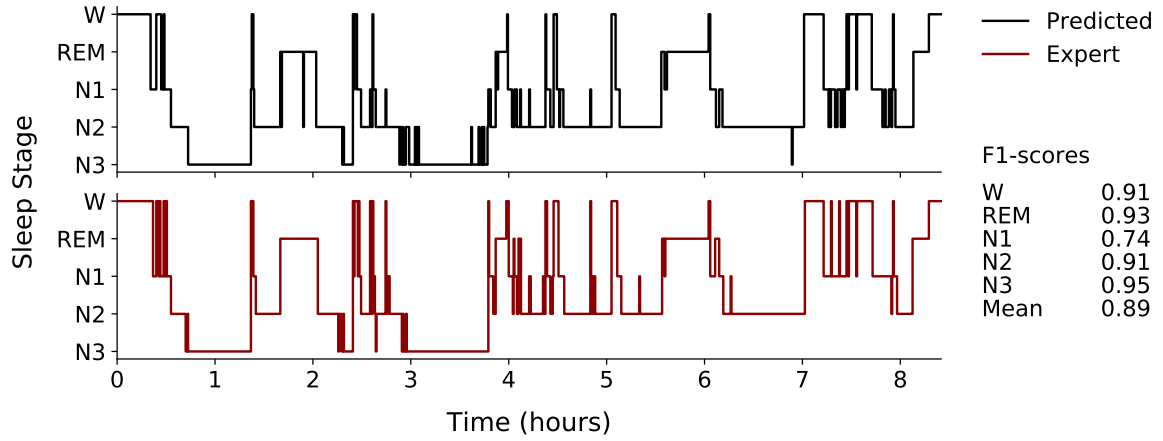


(b) Hypnogram with F1-score nearest dataset median (record `SC4201E0-PSG`).

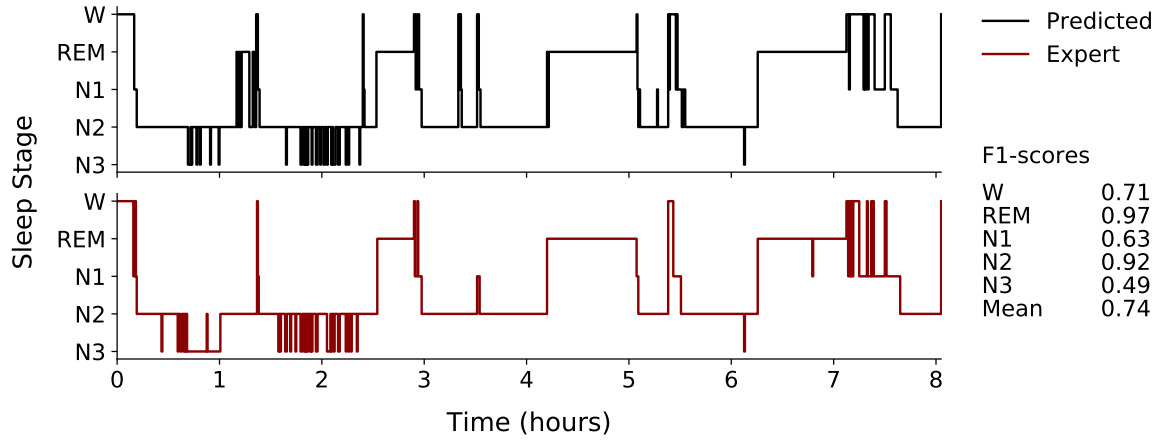


(c) Hypnogram with lowest observed F1-score (record `SC4571F0-PSG`).

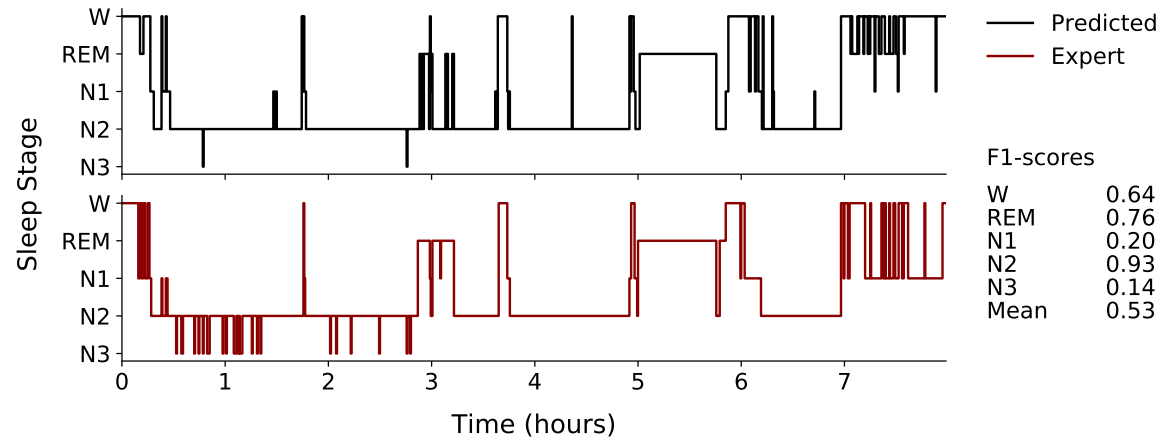
Figure D.12: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `SEDF-SC`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `ST7212J0-PSG`).

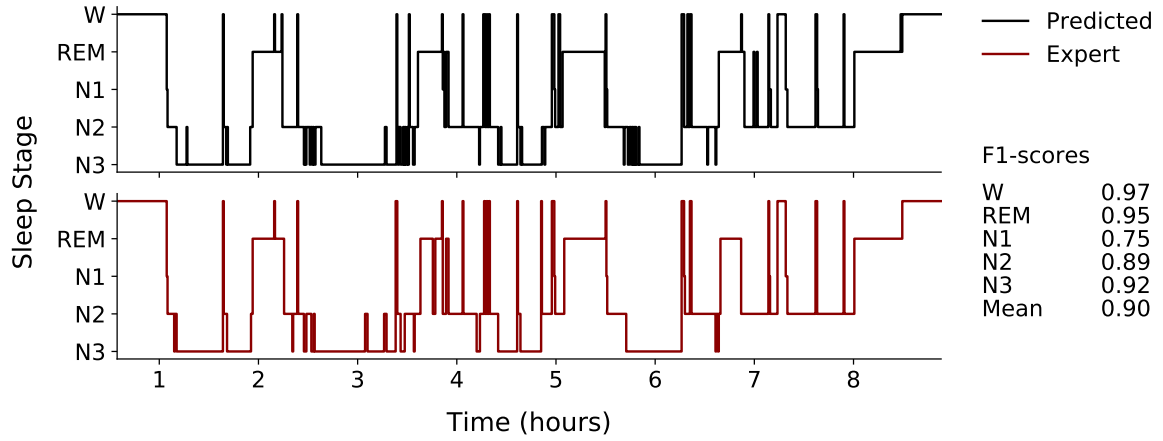


(b) Hypnogram with F1-score nearest dataset median (record `ST7182J0-PSG`).

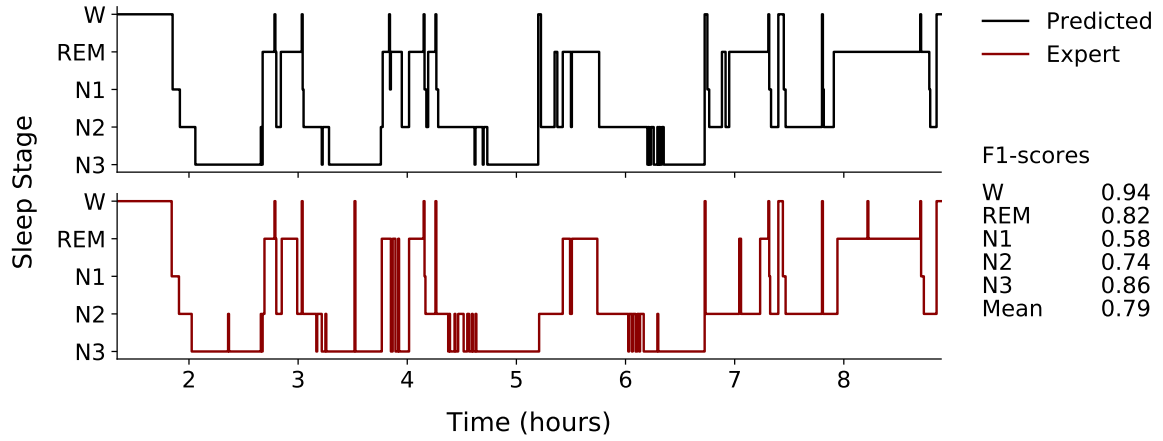


(c) Hypnogram with lowest observed F1-score (record `ST7181J0-PSG`).

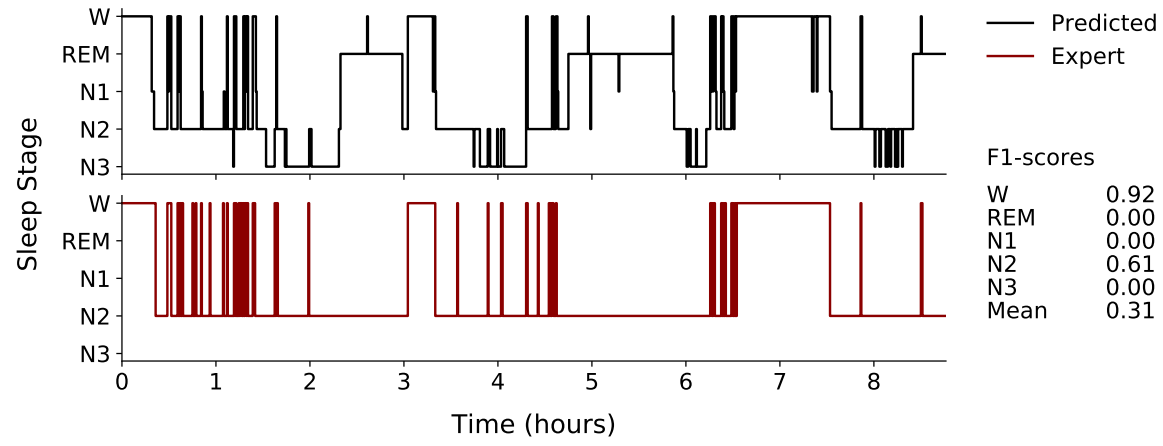
Figure D.13: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `SEDF-ST`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `shhs1-204781`).

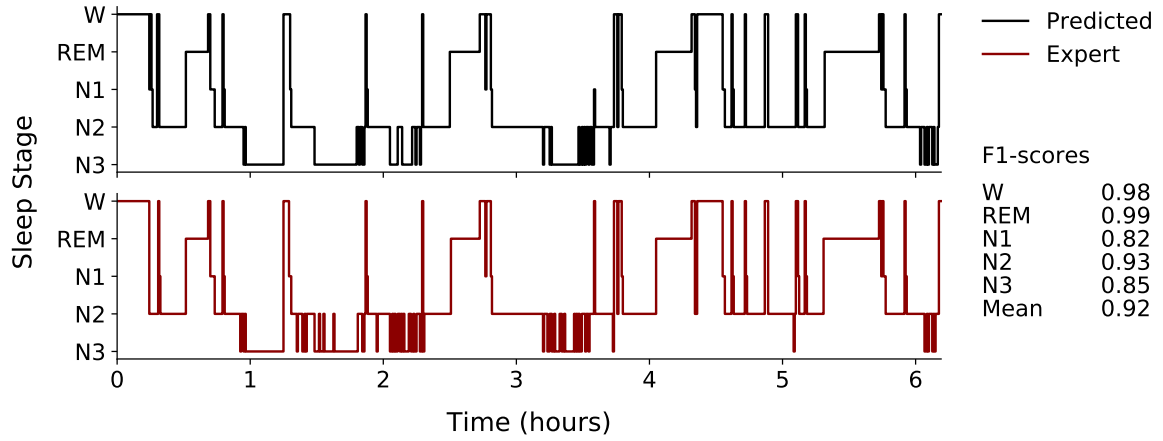


(b) Hypnogram with F1-score nearest dataset median (record `shhs1-204364`).

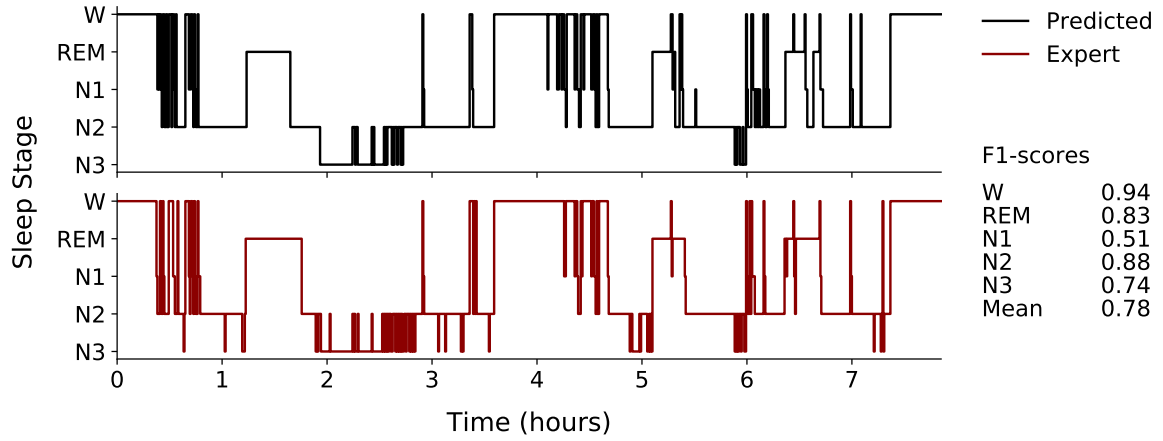


(c) Hypnogram with lowest observed F1-score (record `shhs1-201279`).

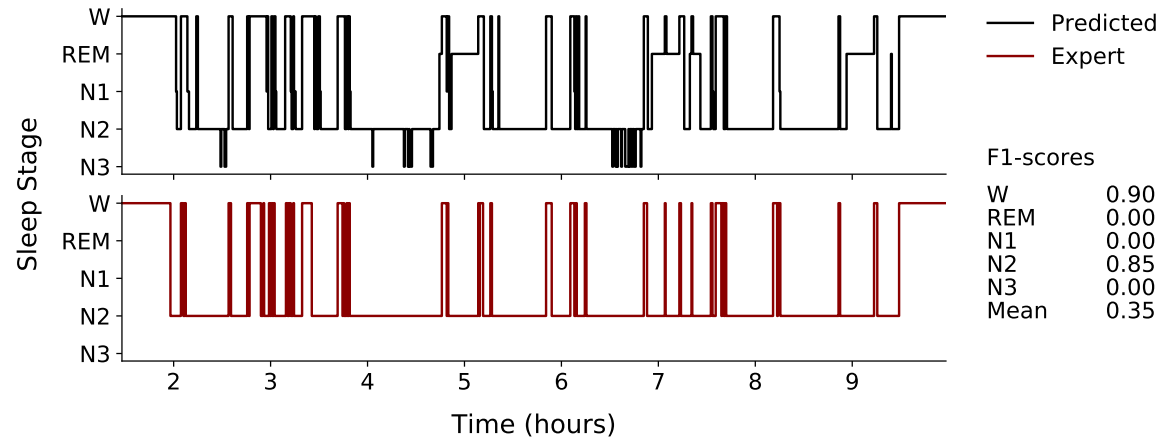
Figure D.14: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `SHHS`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `sof-visit-8-10354`).

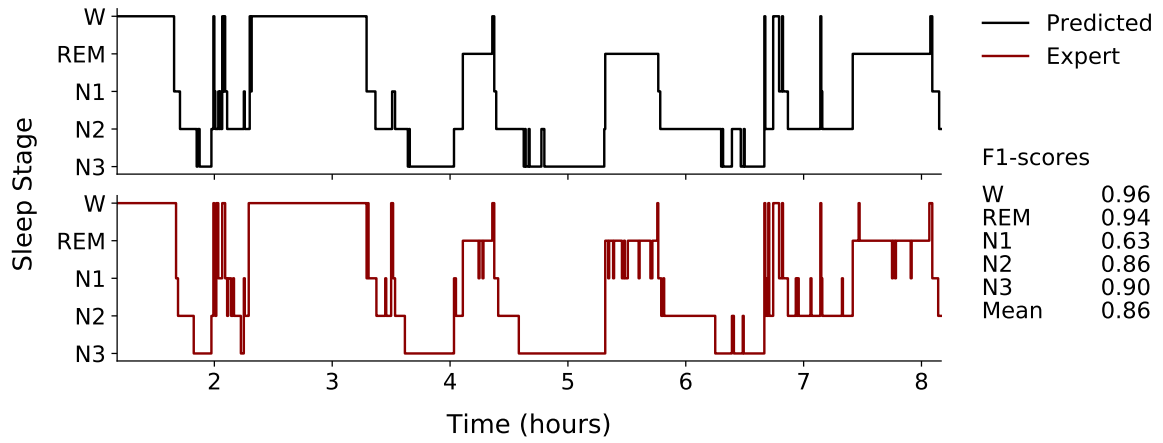


(b) Hypnogram with F1-score nearest dataset median (record `sof-visit-8-09115`).

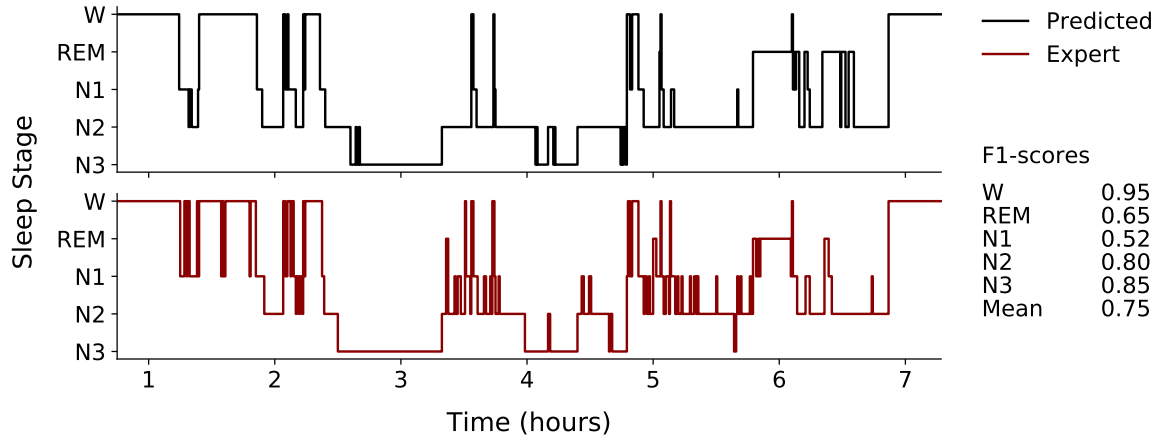


(c) Hypnogram with lowest observed F1-score (record `sof-visit-8-10514`).

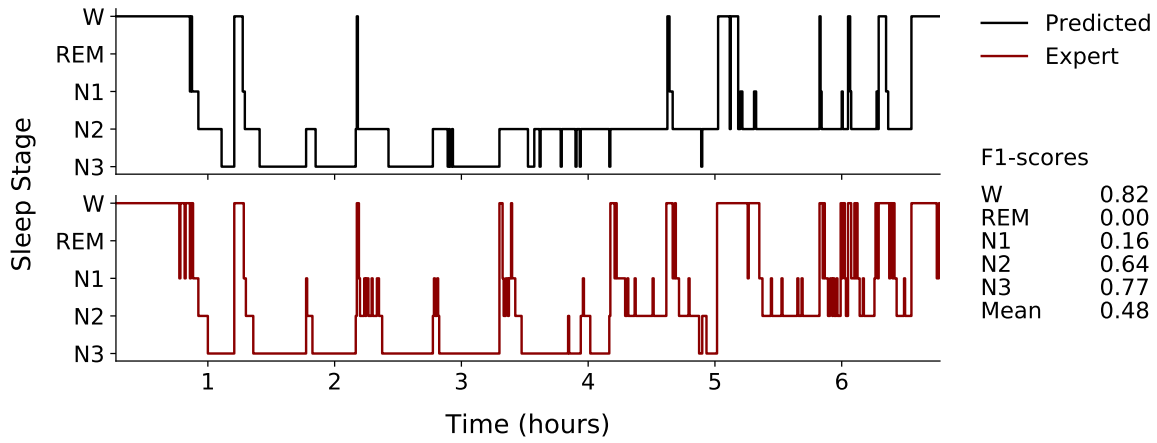
Figure D.15: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `SOF`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `subject_48`).

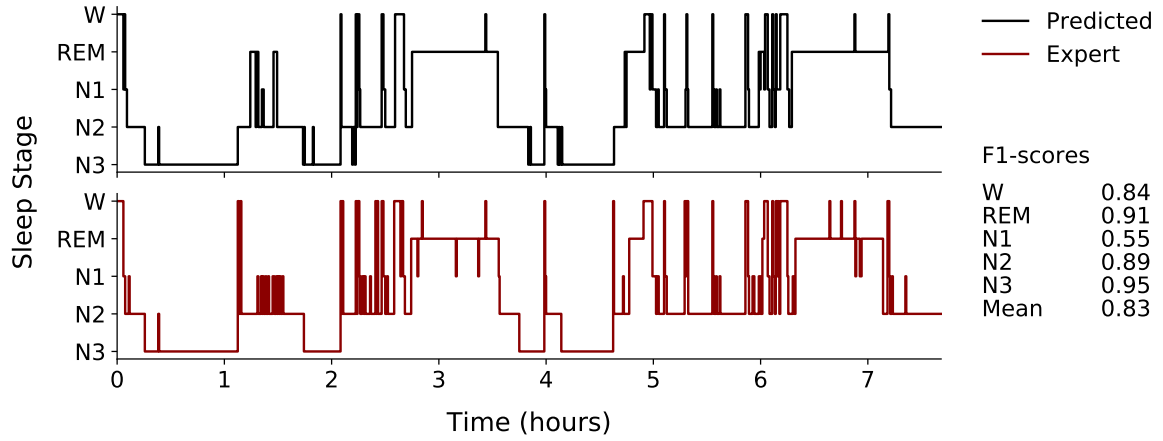


(b) Hypnogram with F1-score nearest dataset median (record `subject_5`).

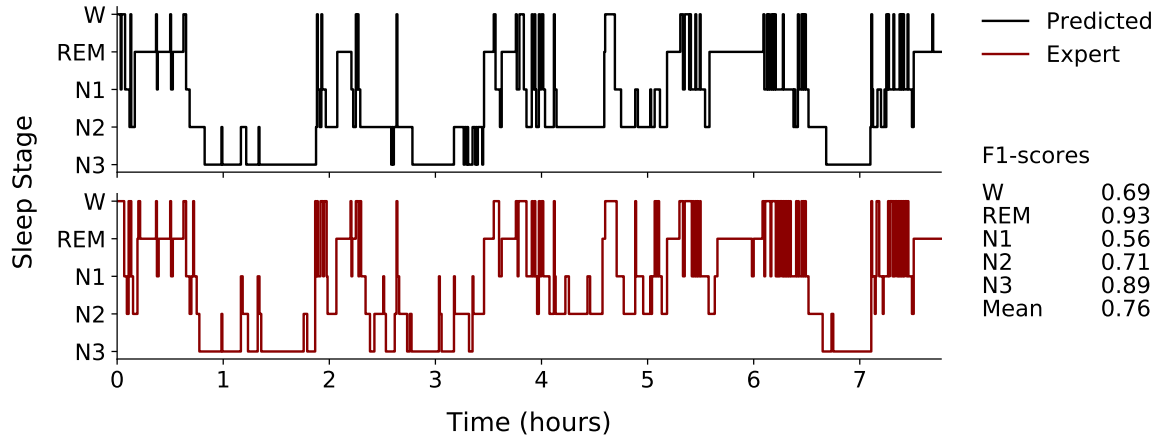


(c) Hypnogram with lowest observed F1-score (record `subject_54`).

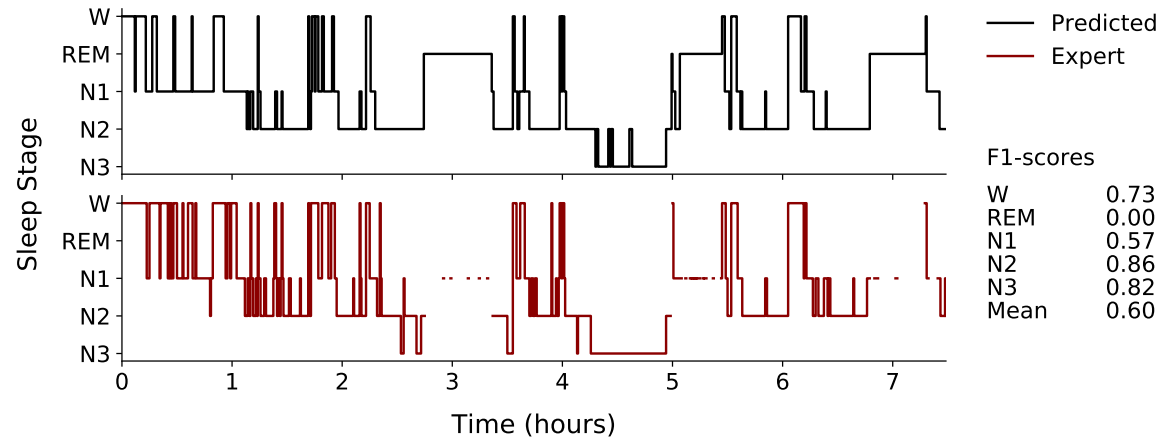
Figure D.16: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `ISRUC-SG1`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `subject_8_visit_2`).

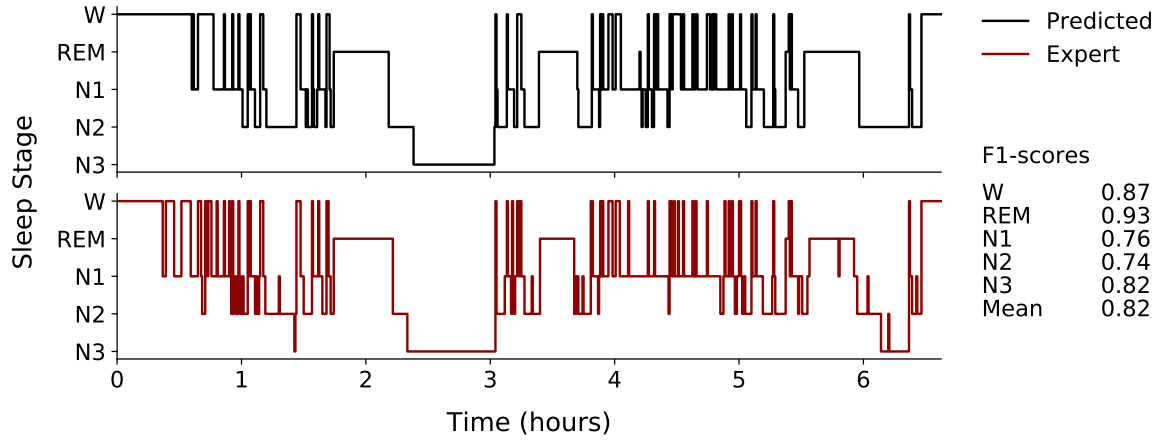


(b) Hypnogram with F1-score nearest dataset median (record `subject_1_visit_1`).

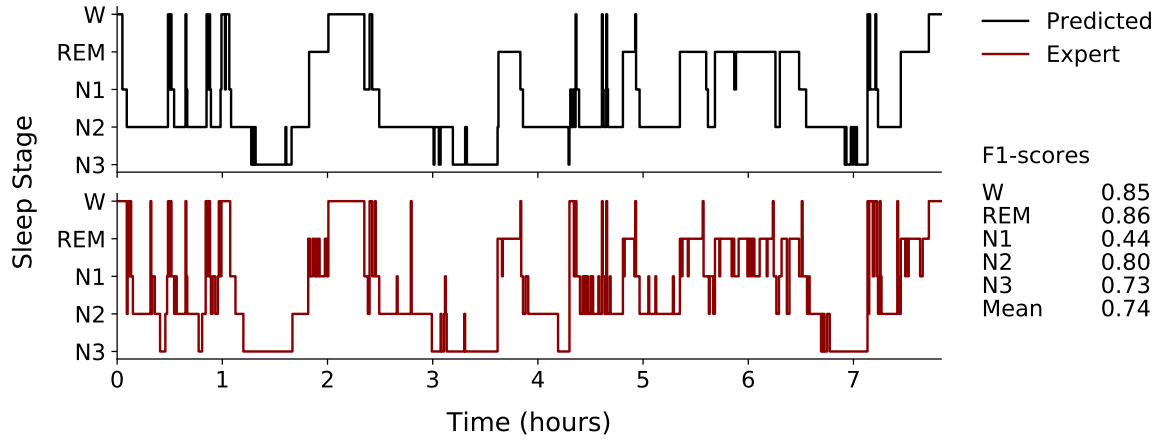


(c) Hypnogram with lowest observed F1-score (record `subject_7_visit_2`).

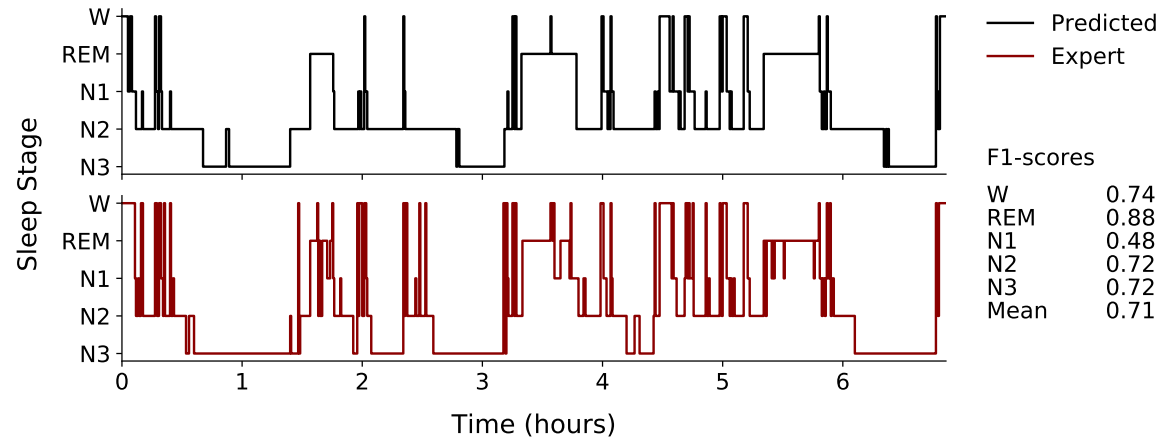
Figure D.17: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `ISRUC-SG2`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `subject_10`).

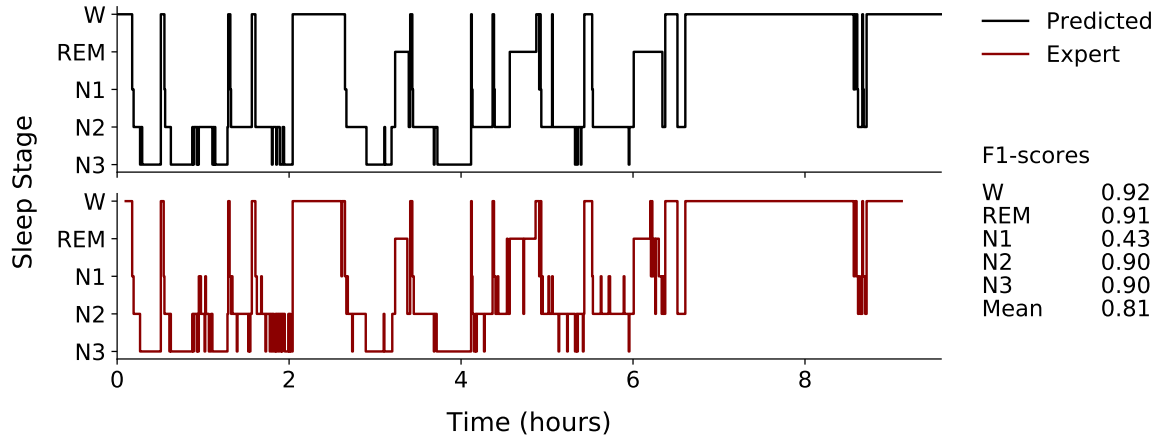


(b) Hypnogram with F1-score nearest dataset median (record `subject_2`).

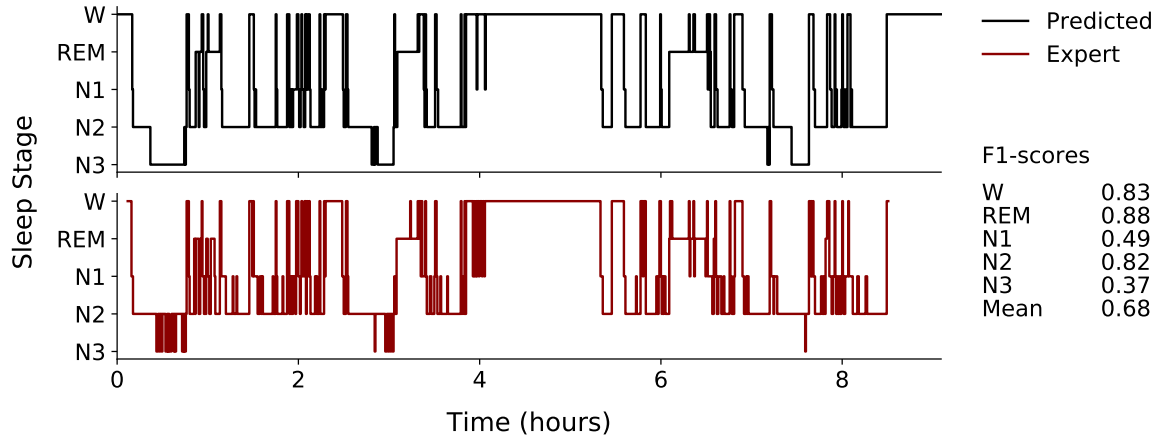


(c) Hypnogram with lowest observed F1-score (record `subject_3`).

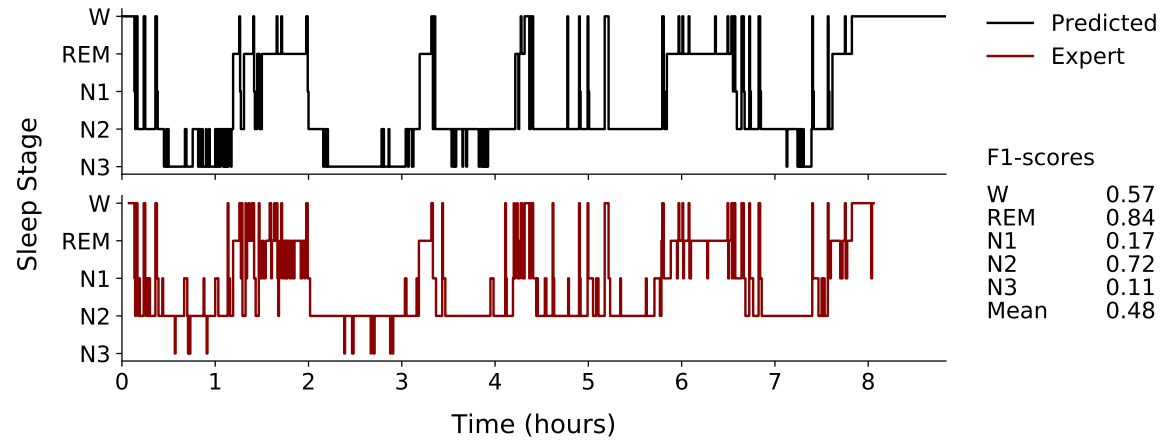
Figure D.18: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `ISRUC-SG3`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `01-01-0016`).

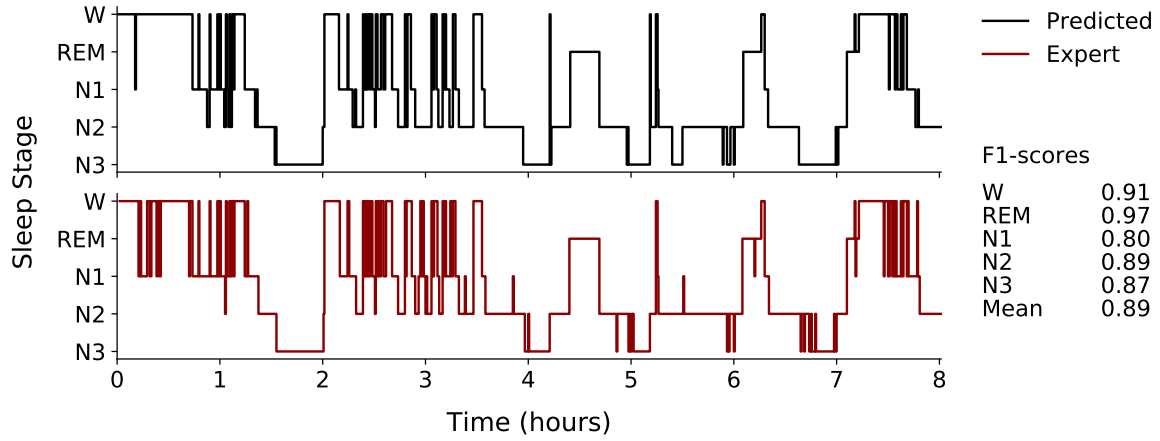


(b) Hypnogram with F1-score nearest dataset median (record `01-01-0047`).

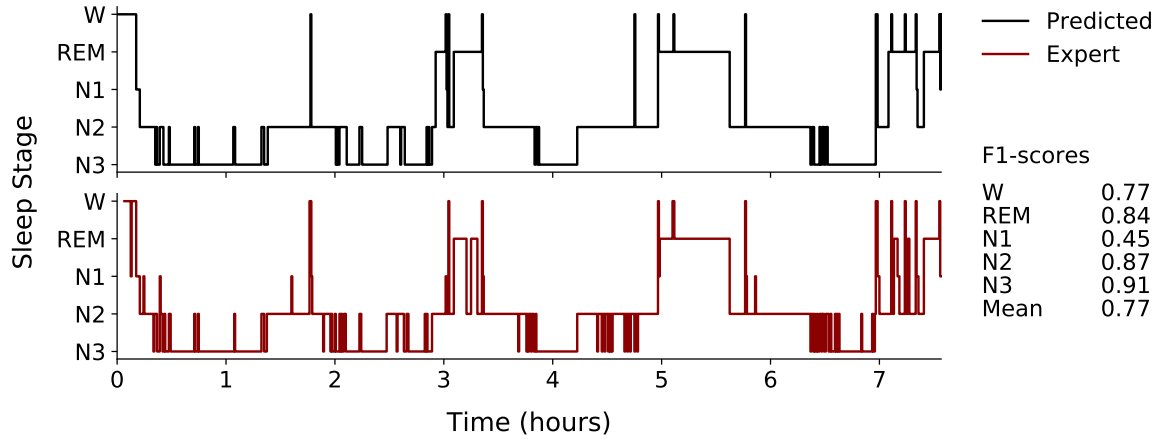


(c) Hypnogram with lowest observed F1-score (record `01-01-0031`).

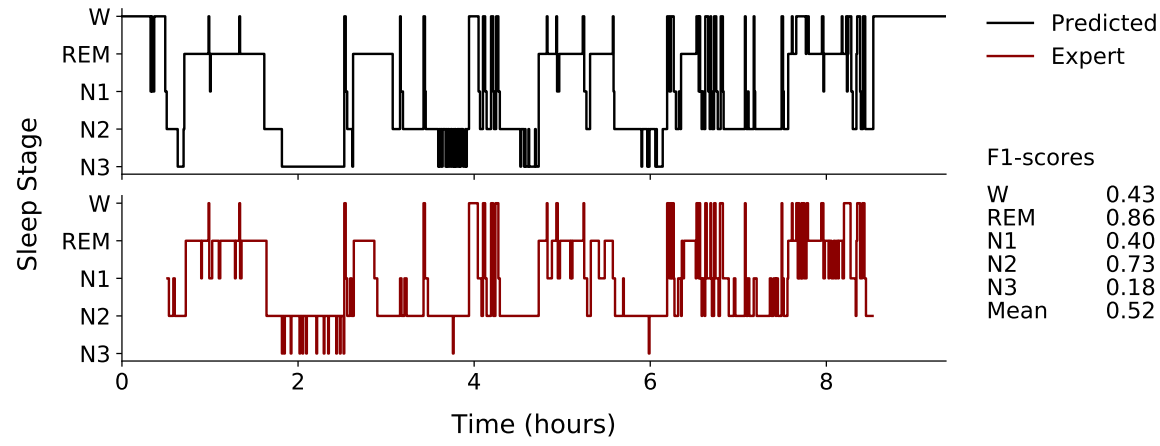
Figure D.19: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `MASS-C1`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `01-03-0035`).

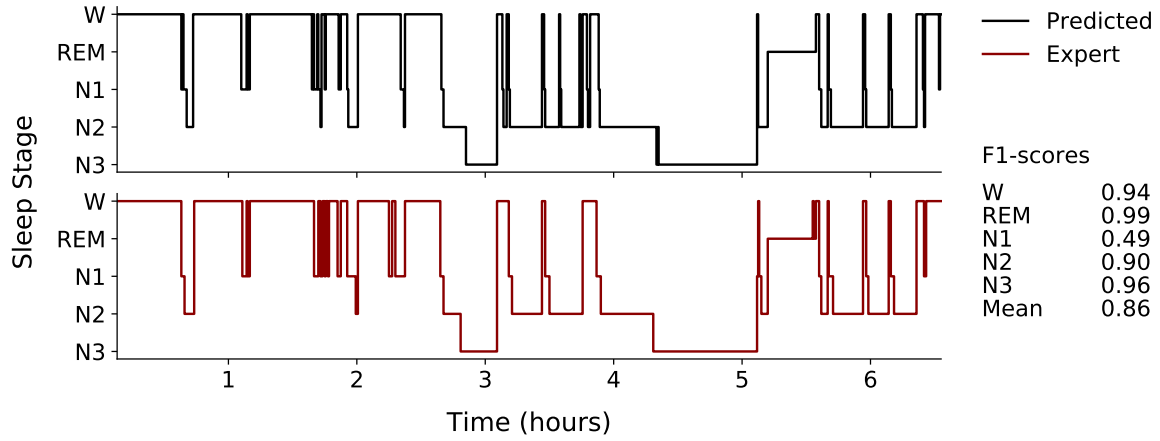


(b) Hypnogram with F1-score nearest dataset median (record `01-03-0008`).

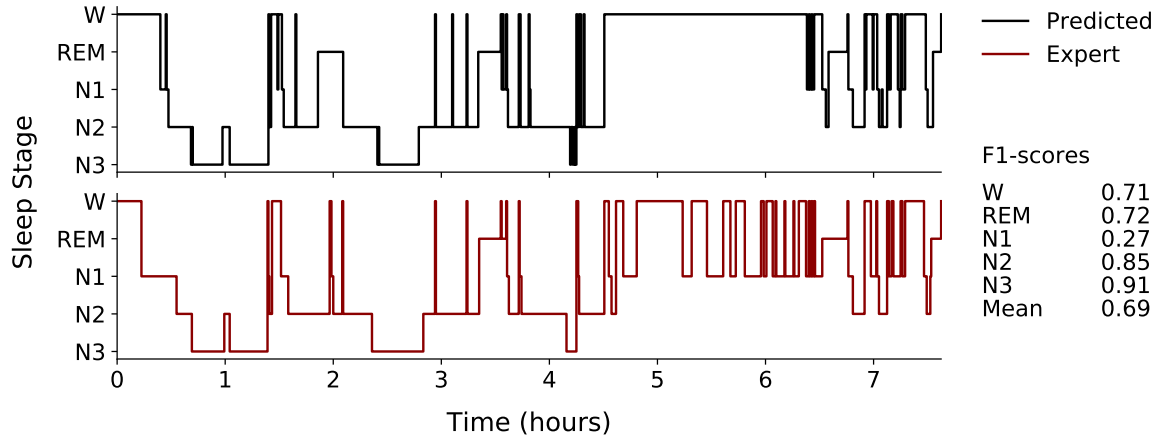


(c) Hypnogram with lowest observed F1-score (record `01-03-0058`).

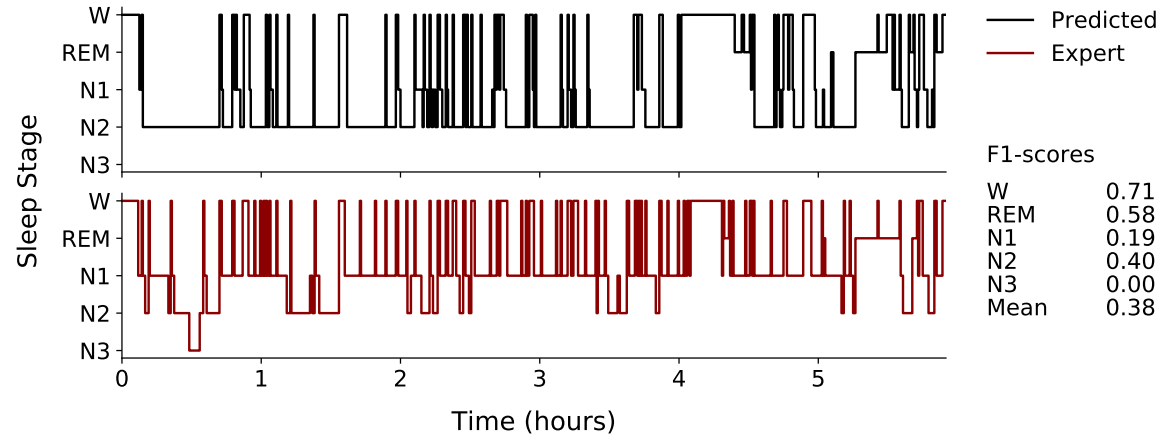
Figure D.20: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `MASS-C3`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `ucddb022`).

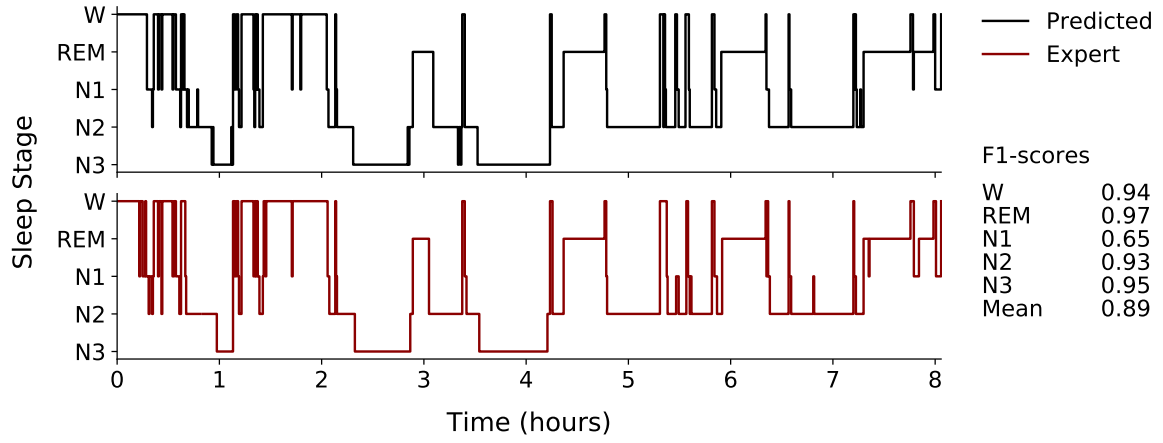


(b) Hypnogram with F1-score nearest dataset median (record `ucddb015`).

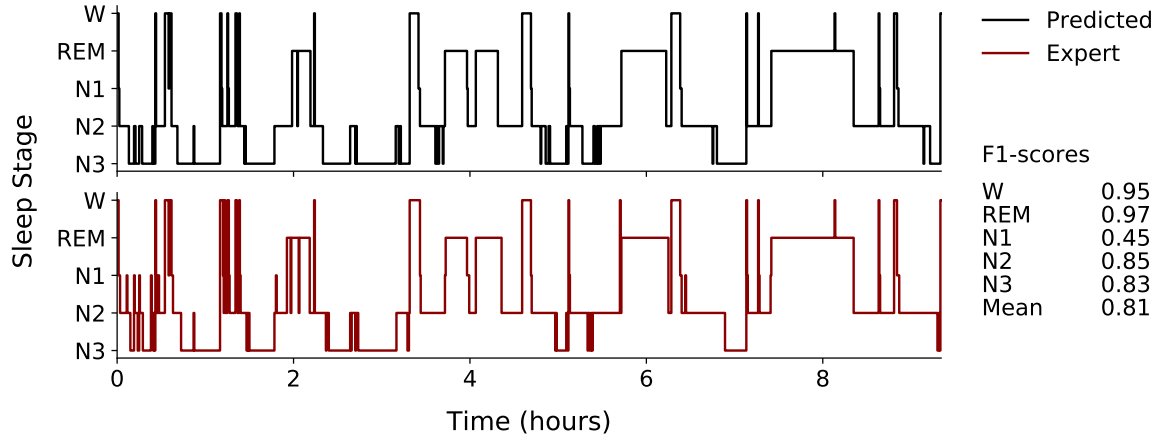


(c) Hypnogram with lowest observed F1-score (record `ucddb025`).

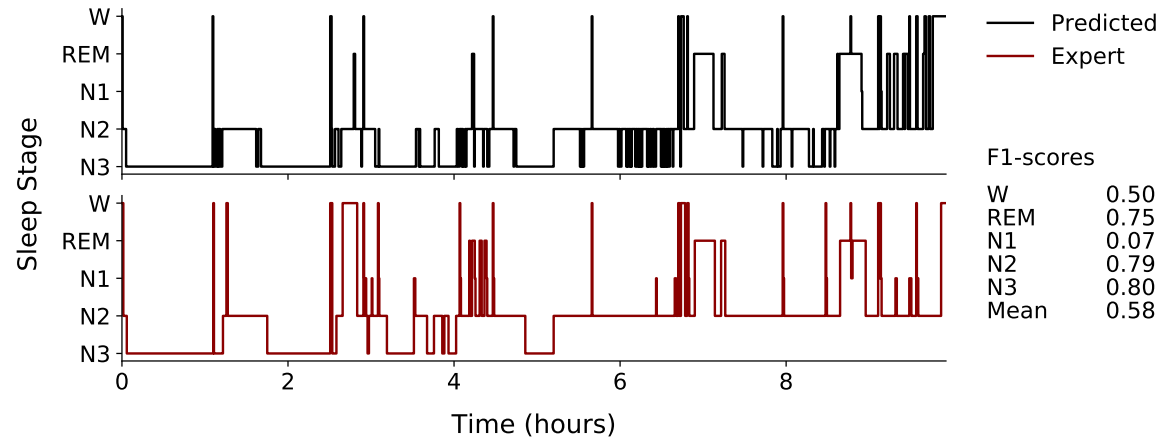
Figure D.21: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `SVUH`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `889dcc46-9998-4b54-9c49-f291f153d101`).

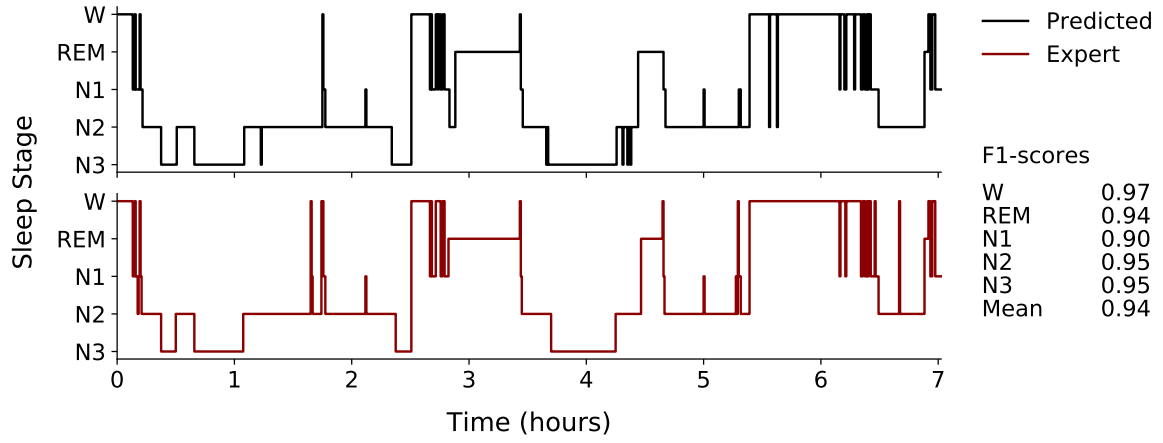


(b) Hypnogram with F1-score nearest dataset median (record `01e60017-d3b5-41cf-bcfb-bde09d46003f`).

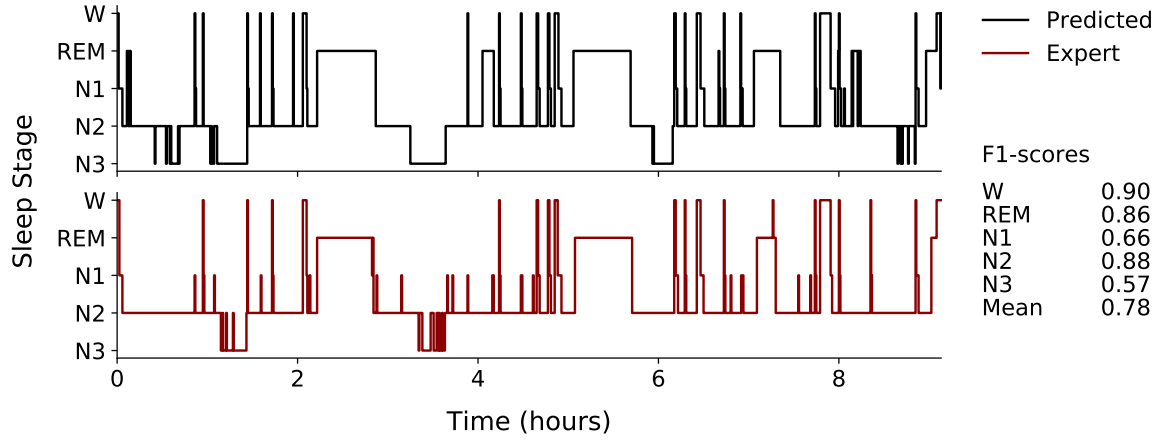


(c) Hypnogram with lowest observed F1-score (record `42f25159-530e-47be-ab07-0895e565ad08`).

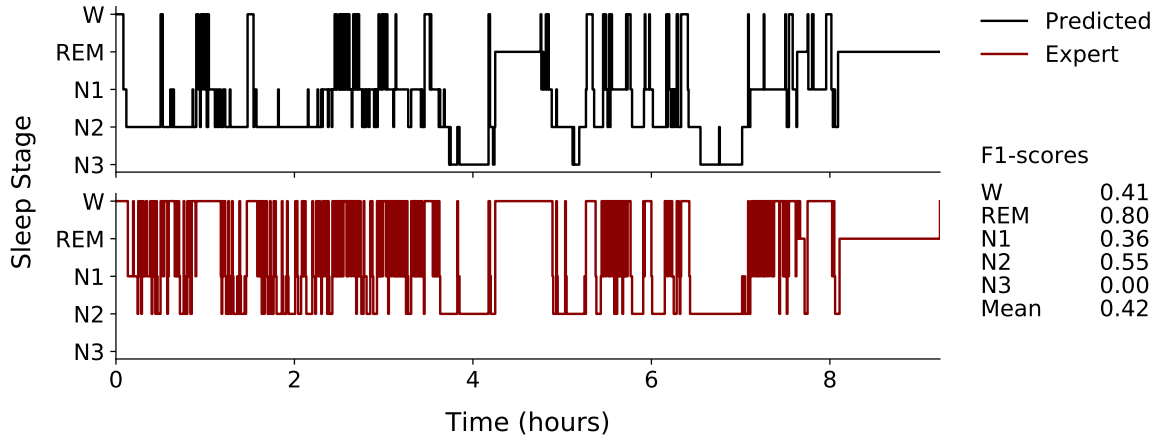
Figure D.22: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `DOD-H`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.



(a) Hypnogram with highest observed F1-score (record `c900fd7f-649d-4ce9-b618-d83c4dea898a`).



(b) Hypnogram with F1-score nearest dataset median (record `7259faa4-fef7-4d76-834d-a1a5e4a04b85`).



(c) Hypnogram with lowest observed F1-score (record `d203e2a0-b261-4b11-9b76-74709094690d`).

Figure D.23: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across records in the test-split of dataset `DOD-O`. Black hypnograms were predicted by U-Sleep, red hypnograms are human expert annotations. F1-scores for each stage are shown to the right of each set of hypnograms. Each hypnogram displays at most 30 minutes of wake prior to and following the first and last non-wake period, respectively, as determined by the human expert annotations.

Table D.3: ABC - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
F4-M1+E2-M1	20	0.81	0.49	0.78	0.60	0.79	0.69
F3-M2+E2-M1	20	0.83	0.48	0.78	0.53	0.85	0.70
C3-M2+E2-M1	20	0.84	0.50	0.81	0.62	0.86	0.73
O1-M2+E2-M1	20	0.84	0.51	0.81	0.63	0.88	0.73
F3-M2+E1-M2	20	0.85	0.48	0.83	0.71	0.90	0.75
O2-M1+E2-M1	20	0.86	0.51	0.83	0.71	0.88	0.76
O1-M2+E1-M2	20	0.85	0.48	0.84	0.73	0.90	0.76
C4-M1+E2-M1	20	0.85	0.51	0.83	0.73	0.87	0.76
F4-M1+E1-M2	20	0.86	0.50	0.83	0.71	0.89	0.76
C3-M2+E1-M2	20	0.86	0.53	0.84	0.75	0.91	0.78
O2-M1+E1-M2	20	0.87	0.52	0.84	0.76	0.90	0.78
C4-M1+E1-M2	20	0.87	0.53	0.84	0.76	0.90	0.78
Mean		0.85	0.50	0.82	0.69	0.88	0.75
Standard deviation		0	0.02	0.02	0.07	0.03	0.03
Majority vote	20	0.87	0.53	0.84	0.72	0.90	0.77

Table D.4: CCSHS - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
C4-A1+ROC-A1	78	0.91	0.57	0.90	0.86	0.90	0.83
C4-A1+LOC-A2	78	0.92	0.57	0.90	0.86	0.91	0.83
C3-A2+ROC-A1	78	0.92	0.58	0.90	0.87	0.91	0.84
C3-A2+LOC-A2	78	0.92	0.59	0.90	0.87	0.91	0.84
Mean		0.92	0.58	0.90	0.86	0.91	0.83
Standard deviation		0	0.01	0.00	0.00	0.00	0.01
Majority vote	78	0.93	0.63	0.91	0.88	0.93	0.85

Table D.5: CFS - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
C3-A2+ROC-A1	92	0.93	0.48	0.87	0.83	0.91	0.80
C3-A2+LOC-A2	92	0.93	0.50	0.87	0.82	0.90	0.81
C4-A1+LOC-A2	92	0.93	0.49	0.88	0.84	0.90	0.81
C4-A1+ROC-A1	92	0.93	0.51	0.88	0.84	0.91	0.81
Mean		0.93	0.50	0.88	0.83	0.91	0.81
Standard deviation		0	0.01	0.00	0.01	0.00	0.00
Majority vote	92	0.93	0.52	0.89	0.84	0.91	0.82

Table D.6: CHAT - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
O2-M1+E1-M2	128	0.91	0.58	0.84	0.89	0.86	0.81
T3-M2+E1-M2	128	0.90	0.60	0.84	0.89	0.87	0.82
T4-M1+E1-M2	128	0.91	0.60	0.84	0.89	0.87	0.82
C4-M1+E1-M2	128	0.92	0.60	0.84	0.89	0.87	0.82
F3-M2+E1-M2	128	0.91	0.60	0.85	0.89	0.88	0.82
C3-M2+E2-M1	128	0.91	0.60	0.85	0.89	0.88	0.83
T3-M2+E2-M1	128	0.90	0.61	0.85	0.89	0.88	0.83
O1-M2+E1-M2	128	0.91	0.60	0.85	0.89	0.88	0.83
T4-M1+E2-M1	128	0.91	0.61	0.85	0.89	0.89	0.83
C3-M2+E1-M2	128	0.91	0.60	0.85	0.90	0.88	0.83
O2-M1+E2-M1	128	0.92	0.60	0.85	0.89	0.88	0.83
F4-M1+E1-M2	128	0.92	0.60	0.85	0.89	0.89	0.83
O1-M2+E2-M1	128	0.91	0.61	0.86	0.89	0.89	0.83
F3-M2+E2-M1	128	0.92	0.62	0.86	0.89	0.89	0.84
F4-M1+E2-M1	128	0.92	0.61	0.86	0.89	0.90	0.84
C4-M1+E2-M1	128	0.93	0.62	0.86	0.89	0.89	0.84
Mean		0.91	0.60	0.85	0.89	0.88	0.83
Standard deviation		0	0.01	0.01	0.00	0.01	0.01
Majority vote	128	0.93	0.64	0.87	0.90	0.90	0.85

Table D.7: DCSM - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
O2-M1+E1-M2	39	0.96	0.47	0.82	0.78	0.85	0.78
O1-M2+E1-M2	39	0.96	0.47	0.83	0.79	0.86	0.78
F3-M2+E1-M2	39	0.97	0.44	0.84	0.81	0.87	0.79
C3-M2+E1-M2	39	0.96	0.45	0.84	0.81	0.87	0.79
O2-M1+E2-M2	39	0.97	0.47	0.84	0.81	0.87	0.79
F4-M1+E1-M2	39	0.97	0.46	0.84	0.80	0.89	0.79
O1-M2+E2-M2	39	0.97	0.48	0.84	0.81	0.87	0.79
F3-M2+E2-M2	39	0.97	0.45	0.84	0.82	0.89	0.79
C4-M1+E1-M2	39	0.97	0.46	0.85	0.80	0.90	0.79
F4-M1+E2-M2	39	0.97	0.46	0.85	0.82	0.88	0.80
C3-M2+E2-M2	39	0.97	0.47	0.85	0.82	0.89	0.80
C4-M1+E2-M2	39	0.97	0.46	0.85	0.82	0.90	0.80
Mean		0.97	0.46	0.84	0.81	0.88	0.79
Standard deviation		0	0.01	0.01	0.01	0.01	0.01
Majority vote	39	0.97	0.48	0.86	0.83	0.89	0.81

Table D.8: HPAP - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
O2-M1+E1	34	0.89	0.44	0.80	0.72	0.86	0.74
C4-M1+E1	34	0.89	0.41	0.80	0.73	0.89	0.75
F4-M1+E2-M1	32	0.89	0.42	0.82	0.74	0.88	0.75
O2-M1+E2-M1	32	0.90	0.43	0.81	0.73	0.88	0.75
F4-M1+E1	34	0.87	0.47	0.80	0.72	0.88	0.75
C3-M2+E1-M2	32	0.90	0.48	0.82	0.71	0.85	0.75
C4-M1+E2-M1	32	0.90	0.43	0.82	0.73	0.88	0.75
C3-M2+E1	34	0.89	0.47	0.82	0.71	0.86	0.75
O2-M1+E2	34	0.89	0.45	0.81	0.75	0.87	0.75
F3-M2+E1	34	0.89	0.46	0.82	0.74	0.87	0.76
C3-M2+E2-M1	32	0.89	0.48	0.83	0.72	0.87	0.76
O1-M2+E1	34	0.89	0.48	0.81	0.73	0.88	0.76
F3-M2+E1-M2	32	0.90	0.47	0.82	0.74	0.87	0.76
F4-M1+E2	34	0.87	0.47	0.81	0.76	0.89	0.76
C4-M1+E2	34	0.89	0.44	0.82	0.75	0.89	0.76
C3-M2+E2	34	0.89	0.47	0.83	0.74	0.86	0.76
F4-M1+E1-M2	32	0.89	0.47	0.82	0.73	0.89	0.76
C4-M1+E1-M2	32	0.90	0.47	0.82	0.73	0.88	0.76
O1-M2+E2-M1	32	0.90	0.47	0.83	0.73	0.88	0.76
F3-M2+E2-M1	32	0.89	0.47	0.83	0.75	0.88	0.76
O2-M1+E1-M2	32	0.90	0.49	0.82	0.73	0.88	0.76
F3-M2+E2	34	0.89	0.47	0.83	0.76	0.87	0.76
O1-M2+E1-M2	32	0.91	0.48	0.83	0.74	0.88	0.77
O1-M2+E2	34	0.89	0.49	0.82	0.76	0.89	0.77
Mean		0.89	0.46	0.82	0.74	0.88	0.76
Standard deviation		0	0.02	0.01	0.01	0.01	0.01
Majority vote	36	0.91	0.48	0.84	0.78	0.90	0.78

Table D.9: MESA - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
Fz-Cz+E2-FPz	100	0.88	0.50	0.83	0.60	0.81	0.72
Fz-Cz+E1-FPz	100	0.90	0.51	0.85	0.66	0.85	0.75
C4-M1+E1-FPz	100	0.91	0.56	0.85	0.62	0.89	0.77
C4-M1+E2-FPz	100	0.91	0.57	0.85	0.65	0.88	0.77
Cz-Oz+E2-FPz	100	0.91	0.58	0.85	0.65	0.88	0.77
Cz-Oz+E1-FPz	100	0.92	0.56	0.86	0.63	0.90	0.77
Mean		0.90	0.55	0.85	0.63	0.87	0.76
Standard deviation		0	0.03	0.01	0.02	0.03	0.02
Majority vote	100	0.92	0.59	0.87	0.65	0.90	0.79

Table D.10: MROS - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
C4-M1+E2-M1	134	0.92	0.44	0.86	0.63	0.87	0.75
C4-M1+E1-M2	134	0.92	0.42	0.86	0.66	0.88	0.75
C3-M2+E1-M2	134	0.92	0.43	0.86	0.68	0.87	0.75
C3-M2+E2-M1	134	0.93	0.45	0.86	0.68	0.86	0.76
Mean		0.92	0.43	0.86	0.66	0.87	0.75
Standard deviation		0	0.01	0.00	0.02	0.01	0.00
Majority vote	134	0.93	0.46	0.87	0.68	0.88	0.77

Table D.11: PHYS - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
O1-M2+E1-M2	100	0.84	0.59	0.82	0.78	0.86	0.78
O2-M1+E1-M2	100	0.84	0.59	0.82	0.79	0.86	0.78
C3-M2+E1-M2	100	0.83	0.59	0.83	0.79	0.87	0.78
F3-M2+E1-M2	100	0.83	0.58	0.83	0.80	0.87	0.78
C4-M1+E1-M2	100	0.83	0.59	0.84	0.80	0.87	0.79
F4-M1+E1-M2	100	0.83	0.59	0.84	0.81	0.87	0.79
Mean		0.83	0.59	0.83	0.79	0.86	0.78
Standard deviation		0	0.00	0.01	0.01	0.01	0.00
Majority vote	100	0.84	0.60	0.84	0.81	0.87	0.79

Table D.12: SEDF-SC - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
Pz-Oz+EOG	23	0.92	0.53	0.83	0.68	0.84	0.76
Fpz-Cz+EOG	23	0.92	0.54	0.86	0.71	0.88	0.78
Mean		0.92	0.54	0.84	0.69	0.86	0.77
Standard deviation		0	0.01	0.01	0.02	0.02	0.01
Majority vote	23	0.93	0.57	0.86	0.71	0.88	0.79

Table D.13: SEDF-ST - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
Fpz-Cz+EOG	8	0.80	0.55	0.87	0.64	0.91	0.76
Pz-Oz+EOG	8	0.79	0.60	0.89	0.63	0.90	0.76
Mean		0.79	0.57	0.88	0.63	0.91	0.76
Standard deviation		0	0.03	0.01	0.01	0.01	0.00
Majority vote	8	0.80	0.58	0.88	0.64	0.91	0.76

Table D.14: SHHS - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
C4-A1+EOG(R)-PG1	140	0.91	0.45	0.86	0.74	0.91	0.77
C3-A2+EOG(L)-PG1	135	0.91	0.49	0.86	0.75	0.89	0.78
C4-A1+EOG(L)-PG1	140	0.92	0.48	0.87	0.76	0.91	0.79
C3-A2+EOG(R)-PG1	135	0.92	0.50	0.87	0.76	0.90	0.79
Mean		0.92	0.48	0.86	0.75	0.90	0.78
Standard deviation		0	0.02	0.00	0.01	0.01	0.01
Majority vote	140	0.93	0.51	0.87	0.76	0.92	0.80

Table D.15: SOF - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
C4-A1+ROC-A1	68	0.93	0.45	0.84	0.70	0.89	0.76
C3-A2+LOC-A2	68	0.91	0.37	0.84	0.78	0.91	0.76
C3-A2+ROC-A1	68	0.92	0.42	0.84	0.74	0.91	0.77
C4-A1+LOC-A2	68	0.93	0.42	0.85	0.76	0.91	0.77
Mean		0.92	0.42	0.84	0.74	0.91	0.77
Standard deviation		0	0.03	0.00	0.03	0.01	0.00
Majority vote	68	0.93	0.45	0.86	0.77	0.92	0.78

Table D.16: ISRUC-SG1 - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
C3-M2+E2-M1	100	0.87	0.50	0.74	0.69	0.82	0.72
C3-M2+E1-M2	100	0.84	0.49	0.77	0.75	0.85	0.74
F3-M2+E2-M1	99	0.88	0.50	0.77	0.72	0.85	0.75
F4-M1+E2-M1	99	0.89	0.50	0.77	0.73	0.85	0.75
O1-M2+E1-M2	100	0.86	0.48	0.78	0.76	0.87	0.75
F4-M1+E1-M2	99	0.86	0.49	0.78	0.76	0.88	0.75
O1-M2+E2-M1	100	0.88	0.49	0.78	0.76	0.86	0.75
O2-M1+E2-M1	100	0.87	0.51	0.77	0.77	0.86	0.76
C4-M1+E2-M1	100	0.88	0.50	0.78	0.77	0.86	0.76
O2-M1+E1-M2	100	0.86	0.48	0.78	0.80	0.87	0.76
F3-M2+E1-M2	99	0.86	0.50	0.79	0.79	0.87	0.76
C4-M1+E1-M2	100	0.86	0.48	0.79	0.81	0.88	0.76
Mean		0.87	0.49	0.77	0.76	0.86	0.75
Standard deviation		0	0.01	0.01	0.03	0.01	0.01
Majority vote	100	0.89	0.52	0.79	0.77	0.88	0.77

Table D.17: ISRUC-SG2 - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
C4-M1+E2-M1	16	0.80	0.48	0.76	0.80	0.84	0.74
C3-M2+E2-M1	16	0.79	0.46	0.77	0.81	0.85	0.74
O1-M2+E2-M1	16	0.80	0.48	0.76	0.82	0.85	0.74
O2-M1+E2-M1	16	0.80	0.48	0.77	0.81	0.85	0.74
F3-M2+E2-M1	16	0.81	0.49	0.77	0.82	0.83	0.74
C4-M1+E1-M2	16	0.84	0.47	0.77	0.82	0.83	0.75
F4-M1+E2-M1	16	0.81	0.49	0.78	0.82	0.86	0.75
F3-M2+E1-M2	16	0.85	0.48	0.78	0.83	0.82	0.75
C3-M2+E1-M2	16	0.85	0.47	0.78	0.83	0.85	0.76
O1-M2+E1-M2	16	0.84	0.47	0.78	0.84	0.86	0.76
O2-M1+E1-M2	16	0.85	0.49	0.78	0.83	0.85	0.76
F4-M1+E1-M2	16	0.87	0.48	0.78	0.83	0.85	0.76
Mean		0.83	0.48	0.77	0.82	0.84	0.75
Standard deviation		0	0.01	0.01	0.01	0.01	0.01
Majority vote	16	0.85	0.49	0.78	0.83	0.86	0.76

Table D.18: ISRUC-SG3 - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
O1-M2+E2-M1	10	0.84	0.54	0.77	0.72	0.87	0.75
C3-M2+E2-M1	10	0.86	0.54	0.77	0.71	0.86	0.75
F4-M1+E1-M2	10	0.91	0.54	0.76	0.69	0.84	0.75
F3-M2+E2-M1	10	0.85	0.53	0.77	0.74	0.85	0.75
O2-M1+E2-M1	10	0.85	0.54	0.77	0.75	0.85	0.75
O1-M2+E1-M2	10	0.91	0.54	0.76	0.70	0.85	0.75
C4-M1+E1-M2	10	0.91	0.49	0.78	0.76	0.84	0.75
C4-M1+E2-M1	10	0.85	0.53	0.78	0.75	0.85	0.75
F4-M1+E2-M1	10	0.86	0.55	0.76	0.72	0.87	0.75
C3-M2+E1-M2	10	0.92	0.53	0.77	0.73	0.85	0.76
O2-M1+E1-M2	10	0.90	0.53	0.79	0.78	0.83	0.76
F3-M2+E1-M2	10	0.92	0.52	0.79	0.78	0.85	0.77
Mean		0.88	0.53	0.77	0.74	0.85	0.75
Standard deviation		0	0.02	0.01	0.03	0.01	0.01
Majority vote	10	0.90	0.55	0.78	0.74	0.85	0.77

Table D.19: MASS-C1 (part 1/2) - Test data - Channel-wise F1/dice scores (computed across subjects). This table displays all MASS-C1 evaluations on EOG(R) channel combinations. Please refer to Table D.20 for MASS-C1 evaluations with EOG(L) channels and for mean, standard deviation and majority vote summaries across both EOG(L) and EOG(R) channel combinations.

	Records	Wake	N1	N2	N3	REM	mean
P3-LER+EOG(R)	6	0.93	0.32	0.66	0.39	0.90	0.64
T4-LER+EOG(R)	6	0.93	0.34	0.68	0.39	0.91	0.65
F8-LER+EOG(R)	6	0.91	0.36	0.72	0.46	0.85	0.66
O2-LER+EOG(R)	6	0.92	0.38	0.72	0.44	0.85	0.66
T6-LER+EOG(R)	6	0.93	0.35	0.71	0.44	0.90	0.66
C4-LER+EOG(R)	6	0.90	0.39	0.73	0.47	0.85	0.67
T3-LER+EOG(R)	6	0.92	0.35	0.73	0.47	0.87	0.67
F4-LER+EOG(R)	6	0.93	0.34	0.72	0.46	0.90	0.67
F3-LER+EOG(R)	6	0.93	0.37	0.72	0.45	0.88	0.67
P4-LER+EOG(R)	6	0.93	0.38	0.73	0.46	0.87	0.67
O1-LER+EOG(R)	6	0.93	0.38	0.73	0.46	0.89	0.68
Pz-LER+EOG(R)	6	0.93	0.40	0.75	0.49	0.85	0.68
C3-LER+EOG(R)	6	0.93	0.37	0.74	0.48	0.89	0.68
P4-CLE+EOG(R)	47	0.93	0.29	0.78	0.57	0.88	0.69
Cz-LER+EOG(R)	6	0.92	0.38	0.75	0.51	0.89	0.69
T5-LER+EOG(R)	6	0.93	0.39	0.75	0.50	0.88	0.69
Fz-LER+EOG(R)	6	0.93	0.40	0.75	0.50	0.90	0.69
F7-LER+EOG(R)	6	0.93	0.39	0.75	0.51	0.90	0.70
Fp1-LER+EOG(R)	3	0.93	0.34	0.80	0.50	0.93	0.70
Cz-CLE+EOG(R)	47	0.93	0.35	0.78	0.57	0.88	0.70
Fp2-LER+EOG(R)	3	0.93	0.39	0.80	0.51	0.91	0.71
T3-CLE+EOG(R)	47	0.93	0.37	0.79	0.58	0.88	0.71
F4-CLE+EOG(R)	47	0.93	0.38	0.79	0.59	0.88	0.71
F7-CLE+EOG(R)	47	0.94	0.40	0.80	0.60	0.87	0.72
P3-CLE+EOG(R)	47	0.93	0.40	0.80	0.60	0.87	0.72
T5-CLE+EOG(R)	47	0.94	0.40	0.80	0.60	0.88	0.72
Fz-CLE+EOG(R)	47	0.94	0.39	0.81	0.61	0.88	0.73
F3-CLE+EOG(R)	47	0.93	0.42	0.80	0.61	0.87	0.73
C3-CLE+EOG(R)	47	0.94	0.40	0.81	0.62	0.88	0.73
O1-CLE+EOG(R)	47	0.94	0.40	0.81	0.62	0.88	0.73
O2-CLE+EOG(R)	47	0.93	0.42	0.82	0.61	0.88	0.73
F8-CLE+EOG(R)	47	0.93	0.41	0.82	0.63	0.88	0.73
C4-CLE+EOG(R)	47	0.93	0.39	0.83	0.65	0.88	0.74
T4-CLE+EOG(R)	47	0.94	0.41	0.83	0.64	0.88	0.74
Pz-CLE+EOG(R)	47	0.94	0.42	0.83	0.64	0.88	0.74
T6-CLE+EOG(R)	47	0.94	0.41	0.83	0.65	0.88	0.74

Table D.20: MASS-C1 (part 2/2) - Test data - Channel-wise F1/dice scores (computed across subjects). This table displays all MASS-C1 evaluations on EOG(L) channel combinations. Please refer to Table D.19 for MASS-C1 evaluations with EOG(R) channels. The displayed mean, standard deviation and majority vote scores represent summaries computed across both EOG(L) (this table) and EOG(R) (Table D.19) channel combinations.

	Records	Wake	N1	N2	N3	REM	mean
F8-LER+EOG(L)	6	0.91	0.38	0.74	0.48	0.82	0.67
P3-LER+EOG(L)	6	0.93	0.36	0.71	0.44	0.90	0.67
O2-LER+EOG(L)	6	0.93	0.37	0.72	0.46	0.87	0.67
C4-LER+EOG(L)	6	0.91	0.40	0.74	0.49	0.84	0.68
Pz-LER+EOG(L)	6	0.92	0.41	0.75	0.49	0.84	0.68
T6-LER+EOG(L)	6	0.92	0.41	0.75	0.49	0.88	0.69
T4-LER+EOG(L)	6	0.93	0.38	0.75	0.48	0.91	0.69
C3-LER+EOG(L)	6	0.91	0.38	0.77	0.53	0.87	0.69
T3-LER+EOG(L)	6	0.93	0.39	0.76	0.50	0.90	0.70
P4-LER+EOG(L)	6	0.92	0.41	0.77	0.52	0.87	0.70
O1-LER+EOG(L)	6	0.93	0.39	0.77	0.51	0.90	0.70
F3-LER+EOG(L)	6	0.93	0.42	0.76	0.51	0.88	0.70
T5-LER+EOG(L)	6	0.92	0.42	0.79	0.56	0.87	0.71
F4-LER+EOG(L)	6	0.93	0.43	0.78	0.53	0.90	0.71
F8-CLE+EOG(L)	47	0.94	0.40	0.78	0.58	0.87	0.71
F7-LER+EOG(L)	6	0.93	0.40	0.79	0.58	0.88	0.72
T3-CLE+EOG(L)	47	0.94	0.43	0.78	0.56	0.88	0.72
T4-CLE+EOG(L)	47	0.94	0.42	0.79	0.57	0.87	0.72
Fz-LER+EOG(L)	6	0.93	0.44	0.79	0.55	0.89	0.72
O1-CLE+EOG(L)	47	0.94	0.41	0.79	0.58	0.88	0.72
T6-CLE+EOG(L)	47	0.94	0.41	0.79	0.59	0.88	0.72
Cz-CLE+EOG(L)	47	0.94	0.45	0.79	0.57	0.87	0.72
Cz-LER+EOG(L)	6	0.93	0.44	0.80	0.58	0.89	0.73
P4-CLE+EOG(L)	47	0.94	0.42	0.80	0.59	0.89	0.73
C4-CLE+EOG(L)	47	0.94	0.43	0.81	0.60	0.87	0.73
O2-CLE+EOG(L)	47	0.94	0.44	0.82	0.59	0.86	0.73
C3-CLE+EOG(L)	47	0.94	0.41	0.81	0.62	0.87	0.73
Fp1-LER+EOG(L)	3	0.94	0.43	0.83	0.56	0.91	0.73
Pz-CLE+EOG(L)	47	0.94	0.42	0.82	0.62	0.87	0.73
P3-CLE+EOG(L)	47	0.93	0.43	0.82	0.63	0.86	0.73
Fz-CLE+EOG(L)	47	0.94	0.43	0.81	0.61	0.87	0.73
T5-CLE+EOG(L)	47	0.94	0.44	0.83	0.63	0.87	0.74
F4-CLE+EOG(L)	47	0.93	0.45	0.83	0.63	0.86	0.74
F3-CLE+EOG(L)	47	0.94	0.47	0.83	0.63	0.85	0.74
F7-CLE+EOG(L)	47	0.94	0.48	0.83	0.63	0.86	0.75
Fp2-LER+EOG(L)	3	0.92	0.48	0.85	0.64	0.88	0.75
Mean [Tables D.19 & D.20]		0.93	0.40	0.78	0.55	0.88	0.71
Standard deviation [Tables D.19 & D.20]		0	0.03	0.04	0.07	0.02	0.03
Majority vote [Tables D.19 & D.20]	53	0.94	0.41	0.81	0.61	0.88	0.73

Table D.21: MASS-C3 - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
C3-LER+EOG(L)	62	0.91	0.50	0.83	0.69	0.90	0.76
O2-LER+EOG(L)	62	0.91	0.50	0.82	0.69	0.90	0.77
C3-LER+EOG(R)	62	0.90	0.51	0.83	0.70	0.89	0.77
O2-LER+EOG(R)	62	0.91	0.51	0.83	0.70	0.90	0.77
F8-LER+EOG(L)	62	0.92	0.50	0.84	0.70	0.91	0.77
F8-LER+EOG(R)	62	0.92	0.51	0.85	0.72	0.90	0.78
T3-LER+EOG(L)	62	0.92	0.50	0.85	0.72	0.91	0.78
F7-LER+EOG(L)	62	0.92	0.52	0.85	0.72	0.91	0.78
Oz-LER+EOG(R)	62	0.90	0.52	0.86	0.75	0.89	0.78
P3-LER+EOG(R)	62	0.89	0.51	0.88	0.78	0.87	0.78
F7-LER+EOG(R)	62	0.92	0.53	0.85	0.73	0.90	0.79
Fp2-LER+EOG(L)	62	0.91	0.53	0.85	0.73	0.90	0.79
Cz-LER+EOG(L)	62	0.92	0.52	0.85	0.73	0.90	0.79
Cz-LER+EOG(R)	62	0.91	0.52	0.86	0.74	0.90	0.79
T3-LER+EOG(R)	62	0.92	0.51	0.86	0.74	0.90	0.79
Oz-LER+EOG(L)	62	0.91	0.53	0.86	0.74	0.90	0.79
Fz-LER+EOG(L)	62	0.92	0.53	0.85	0.73	0.91	0.79
P3-LER+EOG(L)	62	0.89	0.51	0.88	0.77	0.88	0.79
Fz-LER+EOG(R)	62	0.91	0.53	0.85	0.74	0.90	0.79
Fp2-LER+EOG(R)	62	0.91	0.55	0.86	0.74	0.90	0.79
Pz-LER+EOG(L)	62	0.92	0.53	0.86	0.74	0.90	0.79
Pz-LER+EOG(R)	62	0.92	0.53	0.86	0.74	0.89	0.79
T6-LER+EOG(L)	62	0.91	0.52	0.87	0.75	0.90	0.79
T6-LER+EOG(R)	62	0.91	0.53	0.87	0.75	0.90	0.79
T4-LER+EOG(R)	62	0.91	0.55	0.87	0.76	0.88	0.79
F3-LER+EOG(L)	62	0.91	0.53	0.87	0.75	0.90	0.79
F3-LER+EOG(R)	62	0.91	0.54	0.87	0.76	0.90	0.79
T5-LER+EOG(L)	62	0.92	0.53	0.88	0.76	0.90	0.80
O1-LER+EOG(L)	62	0.92	0.53	0.88	0.76	0.90	0.80
T4-LER+EOG(L)	62	0.91	0.55	0.88	0.76	0.89	0.80
C4-LER+EOG(R)	62	0.91	0.56	0.88	0.76	0.88	0.80
F4-LER+EOG(L)	62	0.92	0.54	0.88	0.76	0.90	0.80
T5-LER+EOG(R)	62	0.92	0.54	0.88	0.77	0.89	0.80
F4-LER+EOG(R)	62	0.92	0.55	0.88	0.77	0.89	0.80
P4-LER+EOG(R)	62	0.91	0.56	0.88	0.77	0.89	0.80
C4-LER+EOG(L)	62	0.91	0.55	0.88	0.77	0.89	0.80
O1-LER+EOG(R)	62	0.92	0.54	0.88	0.77	0.90	0.80
P4-LER+EOG(L)	62	0.91	0.55	0.88	0.77	0.89	0.80
Fp1-LER+EOG(R)	62	0.91	0.55	0.88	0.77	0.90	0.80
Fp1-LER+EOG(L)	62	0.91	0.55	0.88	0.76	0.90	0.80
Mean		0.91	0.53	0.86	0.74	0.90	0.79
Standard deviation		0	0.02	0.02	0.02	0.01	0.01
Majority vote	62	0.93	0.54	0.87	0.75	0.91	0.80

Table D.22: SVUH - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
C4-A1+EOG(L)	25	0.78	0.37	0.77	0.66	0.85	0.69
C3-A2+EOG(R)	25	0.80	0.36	0.80	0.79	0.86	0.72
C3-A2+EOG(L)	25	0.79	0.36	0.81	0.81	0.87	0.73
C4-A1+EOG(R)	25	0.80	0.38	0.81	0.81	0.86	0.73
Mean		0.79	0.37	0.80	0.77	0.86	0.72
Standard deviation		0	0.01	0.02	0.06	0.01	0.02
Majority vote	25	0.80	0.37	0.81	0.78	0.88	0.73

Table D.23: DOD-H - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
FP2-F4+EOG2	25	0.83	0.47	0.85	0.79	0.89	0.77
FP2-F4+EOG1	25	0.83	0.48	0.85	0.80	0.89	0.77
FP1-F3+EOG2	25	0.84	0.50	0.85	0.79	0.90	0.78
FP1-F3+EOG1	25	0.83	0.51	0.86	0.80	0.91	0.78
FP1-M2+EOG1	25	0.88	0.56	0.83	0.74	0.92	0.79
FP1-M2+EOG2	25	0.88	0.55	0.83	0.76	0.91	0.79
FP1-O1+EOG1	25	0.88	0.57	0.83	0.76	0.90	0.79
F3-F4+EOG1	25	0.85	0.52	0.86	0.81	0.92	0.79
FP1-O1+EOG2	25	0.89	0.57	0.83	0.76	0.90	0.79
F3-M2+EOG1	25	0.88	0.58	0.83	0.75	0.93	0.79
F3-F4+EOG2	25	0.87	0.51	0.87	0.82	0.90	0.79
FP2-O2+EOG1	25	0.88	0.58	0.84	0.77	0.91	0.79
FP2-M1+EOG1	25	0.88	0.58	0.84	0.76	0.92	0.80
F3-O1+EOG1	25	0.89	0.58	0.84	0.76	0.91	0.80
FP2-M1+EOG2	25	0.89	0.59	0.84	0.76	0.92	0.80
F4-M1+EOG2	25	0.89	0.58	0.84	0.76	0.92	0.80
FP2-O2+EOG2	25	0.89	0.59	0.84	0.76	0.91	0.80
F3-M2+EOG2	25	0.88	0.58	0.85	0.77	0.92	0.80
F4-O2+EOG1	25	0.89	0.58	0.84	0.78	0.91	0.80
F4-M1+EOG1	25	0.88	0.59	0.85	0.77	0.92	0.80
C3-M2+EOG1	25	0.87	0.59	0.85	0.77	0.92	0.80
F4-O2+EOG2	25	0.89	0.59	0.85	0.77	0.91	0.80
F3-O1+EOG2	25	0.90	0.60	0.85	0.78	0.92	0.81
C3-M2+EOG2	25	0.88	0.58	0.86	0.79	0.93	0.81
Mean		0.87	0.56	0.85	0.77	0.91	0.79
Standard deviation		0	0.04	0.01	0.02	0.01	0.01
Majority vote	25	0.91	0.60	0.87	0.79	0.94	0.82

Table D.24: DOD-O - Test data - Channel-wise F1/dice scores (computed across subjects)

	Records	Wake	N1	N2	N3	REM	mean
C3-M2+EOG2	55	0.88	0.51	0.81	0.70	0.88	0.76
F3-M2+EOG2	55	0.88	0.50	0.82	0.69	0.91	0.76
C4-M1+EOG1	55	0.89	0.51	0.83	0.71	0.89	0.77
F3-M2+EOG1	55	0.88	0.50	0.83	0.70	0.92	0.77
C4-M1+EOG2	55	0.87	0.50	0.84	0.71	0.91	0.77
F3-F4+EOG2	55	0.85	0.47	0.86	0.74	0.91	0.77
C3-M2+EOG1	55	0.89	0.51	0.83	0.71	0.90	0.77
F3-O1+EOG2	55	0.87	0.49	0.84	0.73	0.91	0.77
F4-O2+EOG2	55	0.87	0.49	0.84	0.73	0.92	0.77
F4-O2+EOG1	55	0.88	0.50	0.84	0.72	0.91	0.77
F3-F4+EOG1	55	0.87	0.47	0.86	0.74	0.92	0.77
F3-O1+EOG1	55	0.88	0.49	0.85	0.73	0.92	0.77
O1-M2+EOG2	55	0.87	0.49	0.85	0.76	0.89	0.77
O2-M1+EOG1	55	0.88	0.50	0.85	0.76	0.89	0.78
O1-M2+EOG1	55	0.88	0.50	0.86	0.76	0.90	0.78
O2-M1+EOG2	55	0.87	0.50	0.86	0.77	0.90	0.78
Mean		0.88	0.50	0.84	0.73	0.91	0.77
Standard deviation		0	0.01	0.01	0.03	0.01	0.01
Majority vote	55	0.90	0.52	0.86	0.74	0.92	0.79

Appendix F

Appendix for Manuscript F

F.1 U-Sleep v1 on original and corrected datasets

Fiorillo, Monachino, et al. (2023) discovered that the `U-Time` code repository (`https://github.com/perslev/U-Time`) versions `1.0.1` and prior contained a logical error in a function responsible for loading EDF/EDF+ datafiles (resolved as of April 2022) making the original U-Sleep v1 model (Perslev, Darkner, et al. 2021) simultaneously trained on both standard (i.e., as suggested by the AASM guidelines and the International 10-20 system) and non-standard EEG and EOG channel derivations. Interestingly, this effect was not initially discovered because it has minimal impact on model performance. See Fiorillo, Monachino, et al. (2023) and Supplementary Table F.1, which compares U-Sleep v1 on the original dataset and the dataset considered in this present study, which removed the atypical channel derivations.

The unintentional data loading would cause data to be loaded from EDF/EDF+ files where the ordering of returned channels would match that of the loaded file instead of the order requested as per parameters passed to the loading function. This mismatch between the intended and actual function had the following effects on the training of the U-Sleep v1 model:

- For datasets that stored original raw recordings in EDF/EDF+ formatted files, the package `U-Time` either:

    1. Loaded the expected EEG and EOG channels. This occurred for all files where the channel order in the file matched the requested channel order.
    2. Loaded randomly ordered channels (but consistently ordered for all studies in a dataset). For example, two EEGs or an EEG and an EOG might have been loaded incorrectly in the wrong order.
    3. Loaded unexpected and atypical channel derivations such as C3-C4, M1-M2, C3-EOG etc. This sometimes occurred when raw signals were stored without derivations as separate channels (e.g., C3, C4, M1, M2, ...).
    4. A combination of 1 – 3 with some correct channels and some incorrect.

- The U-Sleep v1 model was trained simultaneously on combinations of correctly and incorrectly loaded data, where the incorrectly loaded data itself varied as described above. However, the

set of loaded channels was consistent across all sleep studies within a given dataset that stored identical and similarly ordered channels in their raw files. This was the case for most studies, as most studies within a dataset were collected under similar protocols.

- For studies where channel data were incorrectly loaded (fully or partially), the randomized channel selection routine used to train the U-Sleep v1 model would select random combinations of inputs from these wrongly ordered and/or derived channels in each training iteration.

Consequently, the U-Sleep v1 model was trained on complex and varied input channel modalities and derivations. Although the primary objective of the original training setup was indeed to train the model on diverse and complex inputs to improve robustness and generalization, the intended design at least ensured that the model would always observe exactly 1 EEG and 1 EOG (in that order) of standard derivations according to, for example, the 10-20 system. Instead, the above-described data loading mechanism forced the model to learn to solve the task using typical and atypical inputs.

Importantly, the unintentional data loading did not significantly influence the U-Sleep v1 evaluation results presented in Perslev, Darkner, et al. (2021) because:

- The presented evaluation scores were correct in that they demonstrate the performance of U-Sleep v1 on a wide range of datasets computed across channels that can be obtained using the `utime` software package version `1.0.0` or lower. That is, while some of the input channel types and combinations contributing to the majority scores were atypical from a human expert point of view, they are easily and consistently obtainable in practice and, thus, from a practical and machine learning point of view, equally valid for evaluation as compared to any other set of channels. Supplementary Table F.1 shows a direct comparison of U-Sleep v1 on the original and corrected datasets.

- The most central evaluation, the comparison of U-Sleep v1 with human experts from a previously unseen clinic on the datasets `DOD-H` (healthy controls) and `DOD-O` (sleep-disordered patients), was completely unaffected.

Consequently, all general model performance and robustness statements in Perslev, Darkner, et al. (2021) hold.

The feasibility of this approach is interesting and surprising from a medical sleep point of view. One would expect the type and consistency of input channel modalities and derivations to play a central role in ensuring robust model performance. However, it shows that sleep staging is possible based on simpler representations of brain activity that are detectable from single EEG and EOG channels of atypical and variable derivations and that automatic models can perform sleep staging without

relying on, e.g., AASM guidelines. See Fiorillo, Monachino, et al. (2023) for further discussions and results.

F.1.1 U-Sleep v1 Performance on corrected Data

The U-Sleep v1 model was evaluated on both the *original* dataset and *corrected* dataset. In other words, the U-Sleep v1 model, which was trained on the original dataset containing partly misconfigured channel types and derivations, was applied as-is without further training on the original and corrected testing datasets. These results are shown in Supplementary Table F.1.

All performance metrics computed on the original dataset are identical to those of Perslev, Darkner, et al. (2021). When evaluated on the corrected dataset, some metrics did not change (as rounded to 2 decimal places) either because the original and updated data are identical (i.e., this data was not affected) or because the model performed equally well using either the original or updated data. Considering the median per-subject F1 scores, the U-Sleep v1 model performed significantly worse on stages Wake, N1, N2, REM and Macro F1 ($P < 0.001$ for all) and indifferent on N3 ($P = 0.405$). This is expected because the model was trained on the atypical channel derivations and, therefore, likely to perform slightly better on those exact inputs. However, the absolute median differences were small ($< 0.01 - 0.02$ for all), indicating similar performance and an ability to score using highly variable inputs.

The consensus-scored datasets `DOD-H` and `DOD-O` were unaffected. The performance of U-Sleep v1 in Table 11.2 and Figure 11.1 of the main paper are therefore identical to those listed in Perslev, Darkner, et al. (2021).

F.2 Detailed U-Sleep v2 (EOG) model results

A single-channel EOG model was also developed to support the niche setting where no EEG data is available and to study the feasibility of sleep staging based only on EOG data. Supplementary Table F.3 shows its majority vote performance compared to the U-Sleep v2 model. On average, the single-channel EOG model performed slightly below the U-Sleep v2 model with a weighted mean ± 1 STD macro F1 scores of 0.77 ± 0.04 vs 0.79 ± 0.04 and per-subject median ± 1 MAD scores of 0.75 ± 0.07 vs 0.77 ± 0.07 ($W = 388560$, $P < 0.001$). Similarly, the dual-channel model performed better on all stages Wake, N1, N2, N3 and REM ($P < 0.001$ for all). However, all absolute performance differences were below 0.03 points, and the EOG-only model scored REM nearly as well as the dual-channel model (weighted mean global macro F1 of 0.89 vs 0.90, implying that the EOG-only model was equally able to separate tonic REM stages from Wake.

On the multi-scored `DOD-H` dataset, U-Sleep v2 (EOG) scored worse than U-Sleep v2 ($N = 25$, 0.77 ± 0.10 vs 0.80 ± 0.08, medians 0.81 vs 0.82, $W = 73$, $P = 0.01$) but indifferent from the best

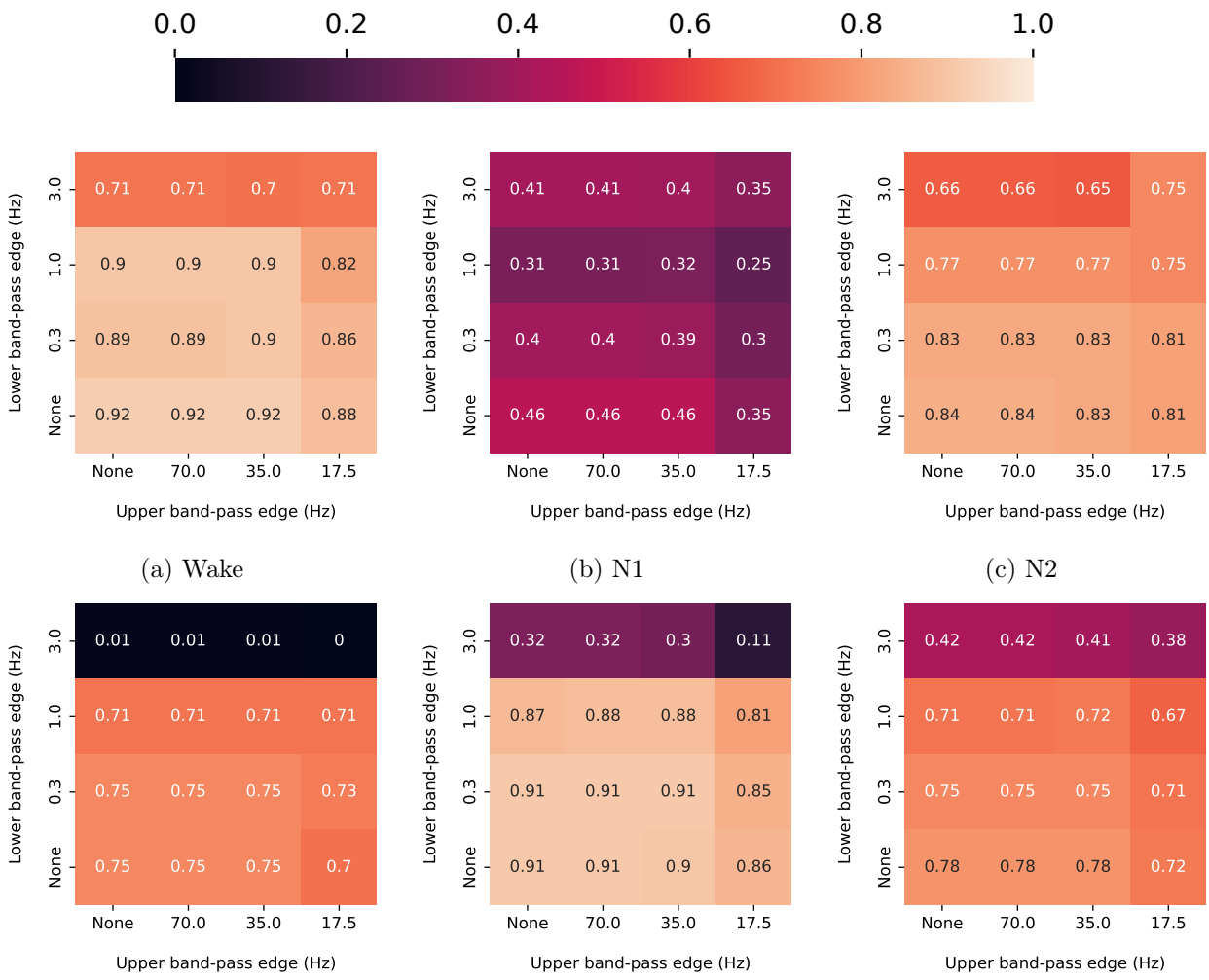
human expert (Expert 3, 0.79 ± 0.06, median 0.80, $W = 140$, $P = 0.56$). On the DOD-O dataset, it scored similarly to U-Sleep v2 ($N = 55$, 0.76 ± 0.10 vs 0.76 ± 0.11, medians 0.79 vs 0.80, $W = 625$, $P = 0.23$) and similar to the best human expert (Expert 5, 0.74 ± 0.11, median 0.77, $W = 597$, $P = 0.15$).

Supplementary Table F.5 shows the performance of U-Sleep v2 (EOG) on the DCSM dataset ($N = 39$ test-split PSGs) when scoring using two individual EOG channels compared to the majority voted hypnogram generated using both channels. The table shows per-subject and per-stage statistics (mean, STD, min, max, median and IQR statistics over F1 scores). As observed for the U-Sleep v2 (EEG) model in the main paper, the majority voted hypnograms are more accurate and should be used when available, but single-channel predictions can produce accurate hypnograms nonetheless. For instance, U-Sleep v2 (EOG) scored with a mean ± 1 STD across macro F1 scores of 0.76 ± 0.10 and median 0.79 using only the { E1-M2 } channel compared to 0.77 ± 0.09, median 0.80 for the majority voted ($N = 2$ channels) hypnograms ($W = 140$, $P < 0.001$). Refer to Supplementary Table F.5 for similar metrics on specific stages and the { E2-M2 } channel.

F.3 Effect of filtering pre-processing

Figure F.1 shows the results of the band-pass filtering experiment. The performance of the U-Sleep v2 model on $N = 36$ test-set sleep studies from the HPAP dataset was evaluated after pre-processing the EEG and EOG input data with various band-pass filtering settings. The mean performance was highest when scoring on raw, un-filtered inputs, with a grand mean F1 score of 0.78. Applying a low-pass filter with an upper bandpass edge of 70.0 Hz or 35.0 Hz had no negative influence on mean performance (0.78 and 0.78, respectively), indicating that U-Sleep v2 did not require information from frequency components higher than 35 Hz to score stages for this dataset. However, applying a low-pass filter at 17.5 Hz significantly reduced mean performance to an F1 score of 0.72. Performance dropped on all five stages, most severely on the N1 stage (from 0.46 to 0.35) and least severely on stage N2 (from 0.84 to 0.81). These results show that frequency components between 17.5 Hz and 35.0 Hz were necessary for optimal scoring of all stages.

High-pass filtering of the data at 0.3 Hz lowered mean performance from 0.78 to 0.75, driven primarily by a drop in N1 and Wake stage accuracy (from 0.46 to 0.40 and from 0.92 to 0.89, respectively). While increasing the lower bandpass edge to 1.0 Hz further decreased mean performance to 0.71, it had no further negative effect on stage Wake performance but led to a significant drop in performance on all stages N1, N2, N3 and REM. As a sanity check, a high-pass filter at 3.0 Hz was also applied, which reduced model accuracy on the slow-wave N3 stage to nearly 0, as the model predicted no N3 stage sleep for most studies in this setting. As expected, this result shows that the



(a) Wake (b) N1 (c) N2

(d) N3 (e) REM (f) Mean

Figure F.1: Filtering experiments. Shown scores are F1/Dice scores computed on summed confusion matrices across all $N = 36$ subjects in the `HPAP` test set.

model relied on slow-wave frequency components in the 0.5 Hz – 2 Hz range as defined by the AASM scoring rules (Iber et al. 2007) to score stage N3.

Applying both a lower- and upper-frequency cut-off had an approximately additive negative influence on overall model performance, and applying the 0.3 Hz – 35 Hz band-pass filter as recommended by the AASM guidelines (Iber et al. 2007) reduced mean performance from the initial F1 score of 0.78 to 0.75. These experiments suggest that raw EEG and EOG inputs may be preferred when using U-Sleep on new data, although the optimal filtering pre-processing may be dataset-specific.

Type	Dataset	Records	Wake	N1	N2	N3	REM	Mean
Internal - Train/Test	ABC	20	0.89	0.53 → 0.42	0.84 → 0.83	**0.72 → 0.73**	**0.90 → 0.91**	0.78 → 0.76
	CCSHS	78	0.97	0.62 → 0.61	0.91	0.88 → 0.87	0.93 → 0.92	0.86
	CFS	92	0.96	0.52	0.88	0.84 → 0.81	0.91 → 0.90	0.82 → 0.81
	CHAT	128	0.96	0.64 → 0.60	0.87 → 0.85	0.90 → 0.88	0.90 → 0.89	0.85 → 0.84
	DCSM	39	0.98	0.47	0.86	0.83	0.88	0.81
	HPAP	36	0.92 → 0.91	0.48 → 0.43	0.84 → 0.83	0.78 → 0.77	0.90	0.78 → 0.77
	MESA	100	0.95	0.59 → 0.46	0.87 → 0.85	**0.65 → 0.72**	0.89	0.79 → 0.77
	MROS	134	0.96 → 0.95	0.46 → 0.44	0.87 → 0.86	**0.68 → 0.69**	0.88 → 0.87	0.77 → 0.76
	PHYS	100	0.84	0.60	0.84	0.81	0.87	0.79
	SEDF-SC	23	0.93	0.57	0.86	0.71	0.88 → 0.87	0.79
	SEDF-ST	8	0.80	0.58 → 0.54	0.88	**0.64 → 0.65**	0.91	0.76
	SHHS	140	0.95 → 0.94	0.51 → 0.50	0.87	**0.76 → 0.77**	0.91	0.80
	SOF	68	0.96	**0.45 → 0.46**	0.86	0.77 → 0.76	0.92	0.79
Hold-Out	ISRUC-SG1	100	0.90 → 0.88	0.52 → 0.50	0.79	**0.77 → 0.78**	**0.88 → 0.89**	0.77
	ISRUC-SG2	16	0.85 → 0.83	**0.49 → 0.50**	0.78	0.83 → 0.82	0.86	0.76
	ISRUC-SG3	10	0.91 → 0.87	**0.55 → 0.56**	0.78	0.74	**0.85 → 0.86**	0.77 → 0.76
	MASS-C1	53	0.94	0.41 → 0.39	0.81	0.61	0.88	0.73
	MASS-C3	62	0.93	0.54 → 0.50	0.87 → 0.85	0.75 → 0.73	0.91	0.80 → 0.78
	SVUH	25	0.82 → 0.81	0.37 → 0.34	0.81	**0.78 → 0.82**	0.88	0.73
	DOD-H	25	0.92	0.60	0.87	0.79	0.94	0.82
	DOD-O	55	0.90	0.52	0.86	0.74	0.92	0.79
	Mean (weighted)		0.93	0.53 → 0.51	0.86 → 0.85	0.77	0.90 → 0.89	0.80 → 0.79
	STD (weighted)		0.04	0.07	0.03	**0.08 → 0.07**	0.02	0.03
	Per subject median		0.95 → 0.94	0.53 → 0.51	0.87 → 0.86	0.80	0.92 → 0.91	0.78
	Per subject MAD		0.03 → 0.04	0.10 → 0.11	0.04	0.11	0.04	0.06
	Pairs w. diff $\neq 0$, n		1076	1063	1090	1007	1018	1092
	Wilcoxon, W		196520	182205	208825	246072	211377	224917
	P-value		< 0.001	< 0.001	< 0.001	0.405	< 0.001	< 0.001

Table F.1: U-Sleep v1 channel-wise majority vote F1 score performance on *original* (left numbers in each cell) and *corrected* (right numbers in each cell) data (see Supplementary Materials section A for details). The layout and statistics of this table mirror that of Table 11.1 in the main paper to which we refer for details.

Table F.2: Channel-wise majority vote sleep staging performance comparisons of U-Sleep v2 (left numbers in each cell) and U-Sleep v2 (EEG) (right numbers in each cell). The layout and statistics of this table mirror that of Table 11.1 in the main paper, to which we refer for details, with the exception that performance on all datasets in this table was included in the summary statistics and test computations.

Type	Dataset	Records	Wake	N1	N2	N3	REM	Macro F1
Internal - Train/Test	ABC	20	0.90	0.62 → 0.53	0.84 → 0.83	0.73 → 0.69	0.92 → 0.91	0.80 → 0.77
	CCSHS	78	0.97	**0.64 → 0.65**	0.92	0.89	0.93	0.87
	CFS	92	0.96	0.53 → 0.52	0.89	0.86 → 0.85	0.92 → 0.91	0.83 → 0.82
	CHAT	128	0.97 → 0.96	0.63 → 0.60	0.87	0.90	0.91	0.85
	DCSM	39	0.99 → 0.97	0.55 → 0.51	0.86 → 0.83	0.83 → 0.82	0.91 → 0.89	0.83 → 0.80
	HPAP	36	**0.92 → 0.93**	0.46 → 0.43	0.84	0.75	0.91 → 0.90	0.78 → 0.77
	MESA	100	**0.95 → 0.96**	0.59 → 0.55	0.87 → 0.86	0.70 → 0.66	0.91 → 0.89	0.80 → 0.78
	MROS	134	0.96	0.44	0.87 → 0.86	0.72 → 0.71	0.89 → 0.88	0.78 → 0.77
	PHYS	100	0.84 → 0.82	0.57 → 0.51	0.85 → 0.84	0.79 → 0.78	0.88	0.79 → 0.77
	SEDF-SC	23	0.93 → 0.92	0.57 → 0.56	**0.86 → 0.87**	**0.76 → 0.80**	0.88 → 0.86	0.80
	SEDF-ST	8	**0.81 → 0.85**	0.60 → 0.53	0.89 → 0.87	0.66 → 0.64	**0.90 → 0.92**	0.77 → 0.76
	SHHS	140	0.94	0.51 → 0.50	0.87	0.77 → 0.73	0.91	0.80 → 0.79
	SOF	68	0.96	**0.45 → 0.47**	0.86	0.77 → 0.75	0.92 → 0.91	0.79
	WSC	218	**0.89 → 0.90**	0.53 → 0.52	0.90	**0.61 → 0.62**	0.90 → 0.89	0.77 → 0.76
	STAGES	89	**0.82 → 0.84**	0.37 → 0.35	0.81	0.69	0.81	0.70
	NCHSDB	102	0.89	0.38 → 0.36	0.86 → 0.85	0.90 → 0.89	0.83	0.77 → 0.76
Hold-Out	ISRUC-SG1	100	**0.90 → 0.91**	0.50 → 0.46	0.78	**0.72 → 0.76**	0.90 → 0.88	0.76
	ISRUC-SG2	16	**0.85 → 0.89**	0.49 → 0.46	**0.76 → 0.77**	**0.74 → 0.77**	0.87 → 0.85	**0.74 → 0.75**
	ISRUC-SG3	10	**0.90 → 0.92**	0.57 → 0.56	**0.74 → 0.75**	**0.62 → 0.64**	**0.86 → 0.87**	**0.74 → 0.75**
	MASS-C1	53	0.94	**0.36 → 0.38**	**0.81 → 0.82**	**0.61 → 0.62**	0.90 → 0.89	**0.72 → 0.73**
	MASS-C3	62	0.93	**0.50 → 0.55**	**0.86 → 0.88**	**0.74 → 0.77**	0.92 → 0.91	**0.79 → 0.81**
	SVUH	25	0.81	**0.29 → 0.33**	0.81 → 0.79	0.86 → 0.74	0.89 → 0.87	0.73 → 0.71
	DOD-H	25	0.90	0.62 → 0.60	**0.88 → 0.92**	**0.82 → 0.89**	0.93	**0.83 → 0.85**
	DOD-O	55	**0.90 → 0.91**	0.51 → 0.48	0.89 → 0.88	**0.78 → 0.79**	0.93 → 0.89	0.80 → 0.79
	DCSM-N	82	0.97	0.48 → 0.45	0.84	**0.80 → 0.81**	0.88	0.80 → 0.79
	DCSM-PLM	41	0.98	0.45 → 0.43	0.84 → 0.83	0.76 → 0.75	0.90	0.79 → 0.78
	DCSM-RBD	34	0.96 → 0.95	0.42 → 0.38	0.82 → 0.80	0.70	0.81 → 0.76	0.74 → 0.72
	DCSM-PD	24	0.95	0.35 → 0.33	**0.67 → 0.70**	**0.60 → 0.66**	**0.65 → 0.67**	**0.65 → 0.66**
	DCSM-RBD-PD	31	0.90 → 0.89	0.29 → 0.28	0.68 → 0.67	**0.55 → 0.61**	0.57 → 0.44	0.60 → 0.58
	Mean (weighted)		0.92	0.50 → 0.48	0.85	0.75	0.89 → 0.87	0.78
	STD (weighted)		**0.05 → 0.04**	**0.09 → 0.08**	0.05	0.09	0.06 → 0.07	0.05
	Per subject median		0.93	0.51 → 0.49	0.87	0.78	0.91 → 0.90	0.77 → 0.76
	Per subject MAD		0.05	0.12	0.04	0.13 → 0.14	0.04 → 0.05	0.07
	Pairs w. diff $\neq 0$, n		1906	1891	1921	1771	1833	1933
	Wilcoxon, W		810105	672535	879098	725634	607209	822175
	P-value		< 0.001	< 0.001	0.071	0.006	< 0.001	< 0.001

Table F.3: Channel-wise majority vote sleep staging performance comparisons of U-Sleep v2 (left numbers in each cell) and U-Sleep v2 (EOG) (right numbers in each cell). The layout and statistics of this table mirror that of Table 11.1 in the main paper, to which we refer for details, with the exception that performance on all datasets in this table was included in the summary statistics and test computations.

Type	Dataset	Records	Wake	N1	N2	N3	REM	Macro F1
Internal - Train/Test	ABC	20	0.90 → 0.88	0.62 → 0.55	0.84 → 0.83	0.73 → 0.69	0.92 → 0.90	0.80 → 0.77
	CCSHS	78	0.97	0.64 → 0.60	0.92 → 0.91	0.89 → 0.88	0.93 → 0.92	0.87 → 0.85
	CFS	92	0.96	0.53 → 0.50	0.89 → 0.88	0.86 → 0.84	0.92 → 0.91	0.83 → 0.82
	CHAT	128	0.97 → 0.96	0.63 → 0.60	0.87 → 0.86	0.90 → 0.89	0.91 → 0.90	0.85 → 0.84
	DCSM	39	0.99 → 0.97	0.55 → 0.49	0.86 → 0.85	0.83 → 0.71	0.91 → 0.89	0.83 → 0.78
	HPAP	36	0.92 → 0.90	0.46	0.84 → 0.81	0.75 → 0.73	0.91 → 0.90	0.78 → 0.76
	MESA	100	0.95	0.59 → 0.54	0.87 → 0.86	0.70 → 0.68	0.91	0.80 → 0.79
	MROS	134	0.96 → 0.95	0.44 → 0.40	0.87 → 0.85	0.72 → 0.65	0.89	0.78 → 0.75
	PHYS	100	0.84 → 0.83	**0.57 → 0.59**	0.85 → 0.83	0.79 → 0.78	0.88 → 0.87	0.79 → 0.78
	SEDF-SC	23	0.93 → 0.90	0.57 → 0.48	0.86 → 0.83	0.76 → 0.66	0.88 → 0.86	0.80 → 0.74
	SEDF-ST	8	0.81 → 0.79	0.60 → 0.53	0.89 → 0.85	0.66 → 0.42	0.90 → 0.89	0.77 → 0.70
	SHHS	140	0.94 → 0.93	0.51 → 0.45	0.87 → 0.86	0.77 → 0.76	0.91	0.80 → 0.78
	SOF	68	0.96	0.45 → 0.39	0.86 → 0.85	0.77 → 0.74	0.92	0.79 → 0.77
	WSC	218	0.89 → 0.87	0.53 → 0.50	0.90 → 0.89	0.61 → 0.57	0.90	0.77 → 0.74
	STAGES	89	**0.82 → 0.83**	0.37 → 0.36	**0.81 → 0.82**	**0.69 → 0.70**	0.81	**0.70 → 0.71**
	NCHSDB	102	0.89	0.38 → 0.36	0.86 → 0.85	0.90 → 0.89	0.83	0.77
Hold-Out	ISRUC-SG1	100	0.90 → 0.84	0.50 → 0.46	0.78 → 0.77	**0.72 → 0.75**	0.90 → 0.88	0.76 → 0.74
	ISRUC-SG2	16	0.85 → 0.79	0.49 → 0.45	0.76	**0.74 → 0.79**	0.87 → 0.86	0.74 → 0.73
	ISRUC-SG3	10	0.90 → 0.86	0.57 → 0.54	**0.74 → 0.77**	**0.62 → 0.74**	0.86 → 0.85	**0.74 → 0.75**
	MASS-C1	53	0.94 → 0.93	**0.36 → 0.43**	**0.81 → 0.82**	0.61	0.90 → 0.89	**0.72 → 0.74**
	MASS-C3	62	0.93 → 0.91	**0.50 → 0.52**	0.86 → 0.84	0.74 → 0.71	0.92 → 0.91	0.79 → 0.78
	SVUH	25	0.81 → 0.79	0.29 → 0.28	0.81 → 0.80	0.86 → 0.75	0.89 → 0.88	0.73 → 0.70
	DOD-H	25	0.90 → 0.83	0.62 → 0.53	0.88 → 0.87	**0.82 → 0.84**	0.93	0.83 → 0.80
	DOD-O	55	0.90 → 0.89	0.51 → 0.48	0.89 → 0.87	0.78 → 0.76	0.93	0.80 → 0.79
	DCSM-N	82	0.97	0.48 → 0.43	0.84	0.80	0.88	0.80 → 0.79
	DCSM-PLM	41	0.98	0.45 → 0.43	0.84 → 0.83	0.76 → 0.74	**0.90 → 0.91**	0.79 → 0.78
	DCSM-RBD	34	0.96 → 0.95	0.42	0.82 → 0.78	0.70 → 0.64	0.81 → 0.77	0.74 → 0.71
	DCSM-PD	24	0.95 → 0.94	0.35 → 0.33	0.67	0.60 → 0.55	0.65 → 0.58	0.65 → 0.62
	DCSM-RBD-PD	31	**0.90 → 0.93**	**0.29 → 0.32**	0.68	0.55 → 0.49	0.57 → 0.53	0.60 → 0.59
	Mean (weighted)		0.92 → 0.91	0.50 → 0.47	0.85 → 0.84	0.75 → 0.73	0.89 → 0.88	0.78 → 0.77
	STD (weighted)		0.05	**0.09 → 0.08**	**0.05 → 0.04**	0.09 → 0.11	0.06	0.05
	Per subject median		0.93 → 0.92	0.51 → 0.47	0.87 → 0.86	0.78 → 0.76	0.91	0.77 → 0.75
	Per subject MAD		0.05 → 0.06	**0.12 → 0.11**	0.04 → 0.05	0.13 → 0.15	0.04	0.07
	Pairs w. diff $\neq 0$, n		1909	1889	1918	1761	1820	1933
	Wilcoxon, W		531146	568069	598905	571044	651190	510951
	P-value		< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

Table F.4: Single-channel per-subject F1 scores for all $N = 39$ test-split PSGs of the `DCSM` dataset scored using the U-Sleep v2 (EEG) model. Each hypnogram was scored using only a single EEG channel at a time (i.e., without the typical majority voting across channels). Statistics shown are mean, standard deviation, minimum and maximum, median (Q_2) and inter-quartile range ($Q_3 - Q_1$) over the per-subject scores. Sub-table (a) shows similar statistics computed on majority-voted hypnograms across all six channels for reference.

(a) *Majority voted ($N = 6$ channels) for comparison*

Stage	Mean	STD	Max	Min	Median	IQR
Wake	0.97	0.12	1.00	0.25	0.99	0.01
N1	0.51	0.13	0.78	0.24	0.53	0.18
N2	0.84	0.12	0.95	0.33	0.88	0.09
N3	0.71	0.29	1.00	0.00	0.83	0.27
REM	0.87	0.13	0.97	0.36	0.91	0.09
Macro	0.78	0.09	0.93	0.54	0.82	0.11

(b) C3-M2

Stage	Mean	STD	Max	Min	Median	IQR
Wake	0.96	0.13	1.00	0.19	0.98	0.01
N1	0.49	0.13	0.71	0.20	0.52	0.16
N2	0.84	0.12	0.94	0.31	0.88	0.06
N3	0.68	0.30	1.00	0.00	0.82	0.28
REM	0.85	0.17	0.97	0.30	0.89	0.08
Macro	0.76	0.11	0.90	0.34	0.80	0.11

(c) C4-M1

Stage	Mean	STD	Max	Min	Median	IQR
Wake	0.98	0.01	1.00	0.92	0.98	0.01
N1	0.50	0.13	0.73	0.24	0.49	0.18
N2	0.84	0.09	0.94	0.58	0.87	0.08
N3	0.68	0.29	1.00	0.00	0.79	0.26
REM	0.87	0.13	0.98	0.38	0.90	0.08
Macro	0.78	0.08	0.90	0.58	0.80	0.12

(d) F3-M2

Stage	Mean	STD	Max	Min	Median	IQR
Wake	0.96	0.13	1.00	0.15	0.99	0.02
N1	0.47	0.14	0.75	0.15	0.49	0.19
N2	0.83	0.12	0.94	0.32	0.87	0.11
N3	0.68	0.30	0.97	0.00	0.84	0.28
REM	0.85	0.17	0.97	0.13	0.91	0.10
Macro	0.76	0.11	0.92	0.38	0.79	0.13

(e) F1-M2

Stage	Mean	STD	Max	Min	Median	IQR
Wake	0.96	0.11	1.00	0.28	0.99	0.02
N1	0.48	0.14	0.68	0.15	0.49	0.14
N2	0.83	0.12	0.94	0.33	0.87	0.12
N3	0.68	0.30	0.96	0.00	0.83	0.26
REM	0.86	0.13	0.99	0.41	0.90	0.10
Macro	0.76	0.08	0.90	0.58	0.78	0.12

(f) O1-M2

Stage	Mean	STD	Max	Min	Median	IQR
Wake	0.96	0.14	1.00	0.14	0.99	0.02
N1	0.48	0.13	0.74	0.19	0.50	0.17
N2	0.80	0.12	0.93	0.33	0.83	0.13
N3	0.66	0.30	1.00	0.00	0.79	0.39
REM	0.80	0.19	0.97	0.03	0.87	0.11
Macro	0.74	0.11	0.91	0.42	0.77	0.14

(g) O2-M1

Stage	Mean	STD	Max	Min	Median	IQR
Wake	0.96	0.13	1.00	0.15	0.99	0.02
N1	0.47	0.14	0.73	0.19	0.50	0.21
N2	0.80	0.12	0.94	0.29	0.81	0.14
N3	0.66	0.30	1.00	0.00	0.78	0.39
REM	0.81	0.19	0.97	0.17	0.89	0.14
Macro	0.74	0.12	0.92	0.26	0.76	0.14

Table F.5: Single-channel per-subject F1 scores for all $N = 39$ test-split PSGs of the `DCSM` dataset scored using the U-Sleep v2 (EOG) model. Each hypnogram was scored using only a single EOG channel at a time (i.e., without the typical majority voting across channels). Statistics shown are mean, standard deviation, minimum and maximum, median (Q_2) and inter-quartile range ($Q_3 - Q_1$) over the per-subject scores. Sub-table (a) shows similar statistics computed on majority-voted hypnograms across all six channels for reference.

(a) *Majority voted ($N = 6$ channels) for comparison*

Stage	Mean	STD	Max	Min	Median	IQR
Wake	0.96	0.12	1.00	0.21	0.98	0.02
N1	0.50	0.11	0.68	0.21	0.53	0.18
N2	0.85	0.08	0.95	0.59	0.87	0.08
N3	0.67	0.29	1.00	0.00	0.81	0.32
REM	0.88	0.12	0.99	0.42	0.92	0.07
Macro	0.77	0.09	0.90	0.52	0.80	0.10

(b) E1-M2

Stage	Mean	STD	Max	Min	Median	IQR
Wake	0.96	0.13	1.00	0.20	0.98	0.02
N1	0.47	0.11	0.65	0.14	0.50	0.15
N2	0.84	0.09	0.95	0.59	0.86	0.09
N3	0.64	0.32	1.00	0.00	0.81	0.38
REM	0.87	0.14	0.99	0.35	0.92	0.07
Macro	0.76	0.10	0.88	0.51	0.79	0.15

(c) E2-M2

Stage	Mean	STD	Max	Min	Median	IQR
Wake	0.96	0.08	1.00	0.50	0.98	0.02
N1	0.50	0.12	0.68	0.23	0.53	0.21
N2	0.84	0.08	0.94	0.57	0.85	0.08
N3	0.65	0.29	1.00	0.00	0.78	0.40
REM	0.87	0.11	0.98	0.57	0.91	0.10
Macro	0.76	0.08	0.89	0.57	0.78	0.11

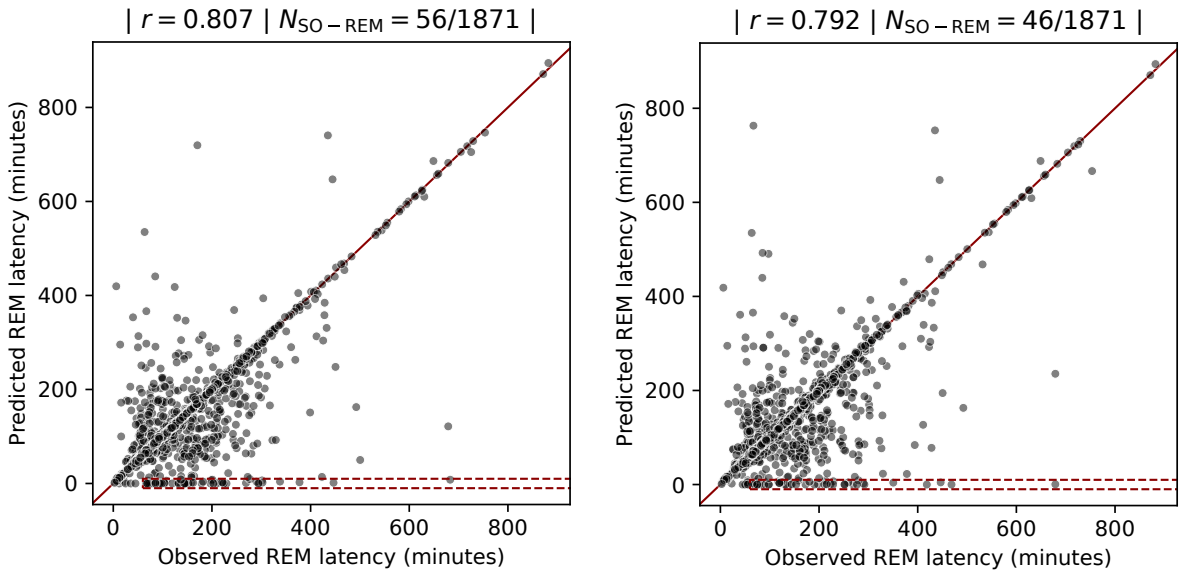
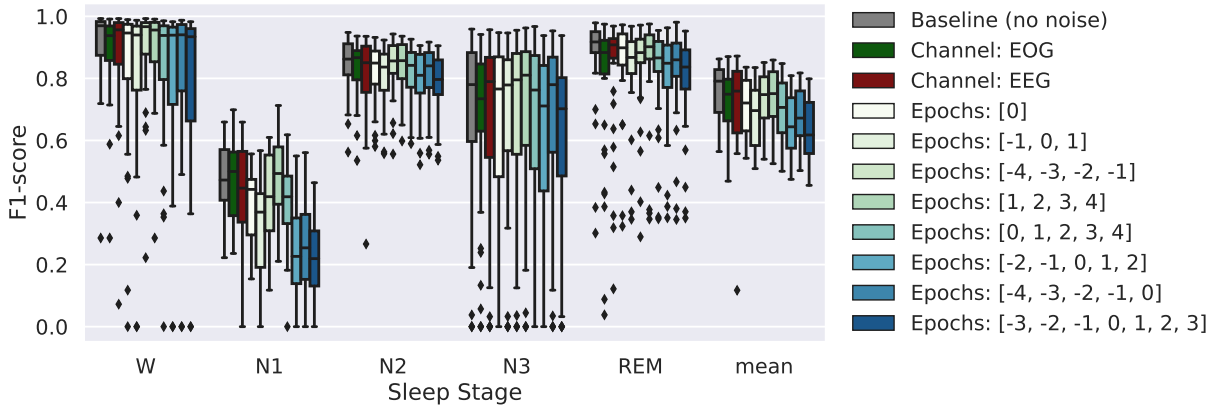
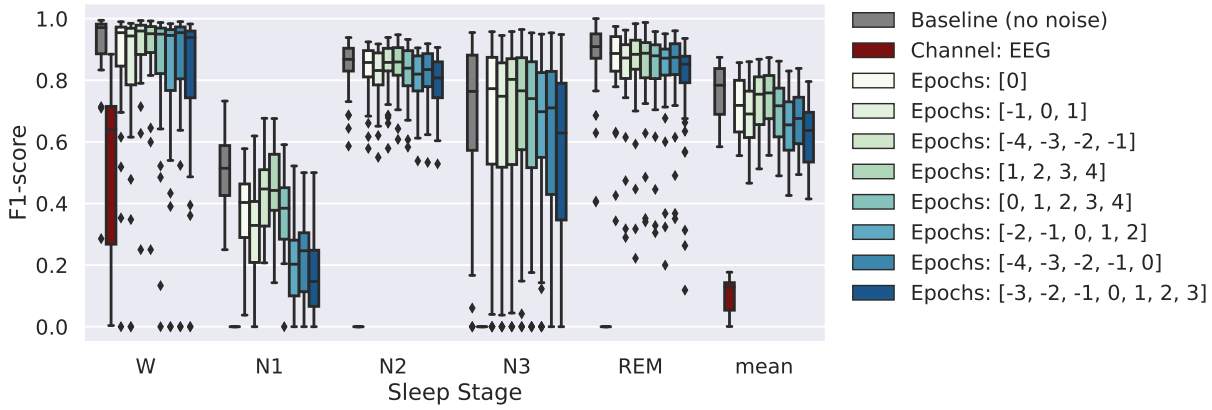


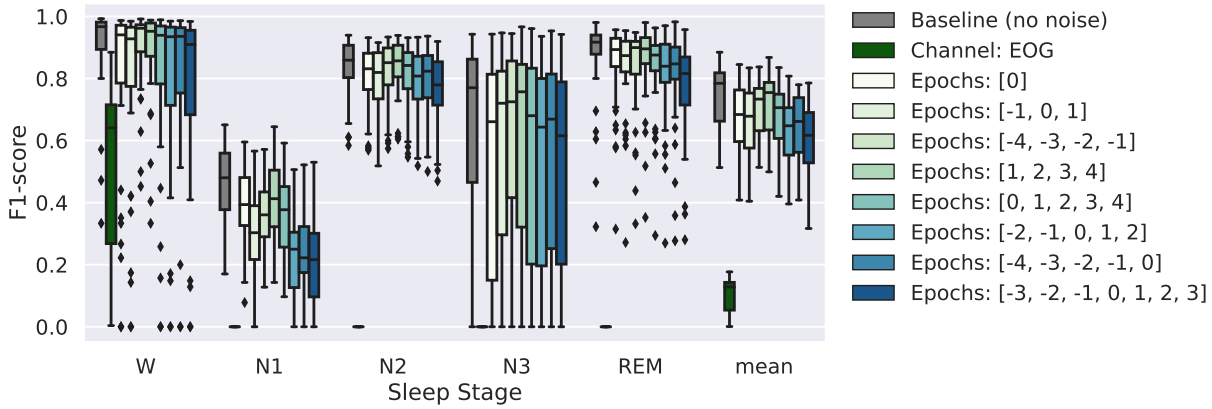
Figure F.2: REM latencies computed from predicted majority-voted hypnograms of U-Sleep v2 (EEG) (left) and U-Sleep v2 (EOG) (right) correlated against observed REM latencies computed from expert annotated hypnograms. Pearson's correlation coefficient, r, is indicated above each plot, along with the number of likely wrongly predicted SO-REMs (dots within the lower red box of each plot; here defined as observations where the observed REM latency was at least 60 minutes while the predicted latency was at most 10 minutes.). Both U-Sleep v2 (EEG) and U-Sleep v2 (EOG) performed REM latency estimation less accurately than the U-Sleep v2 with a lower general correlation and more wrongly predicted SO-REMs (see Figure 11.2.



(a) U-Sleep v1 (EEG + EOG)



(b) U-Sleep v2 (EEG)



(c) U-Sleep v2 (EOG)

Figure F.3: Context predicitons on $N = 39$ test-split sleep studies from the `DCSM` dataset for models U-Sleep v1, U-Sleep v2 (EEG) and U-Sleep v2 (EOG). See also Figure 11.5 for a similar plot for the U-Sleep v2 model and additional methodological details.

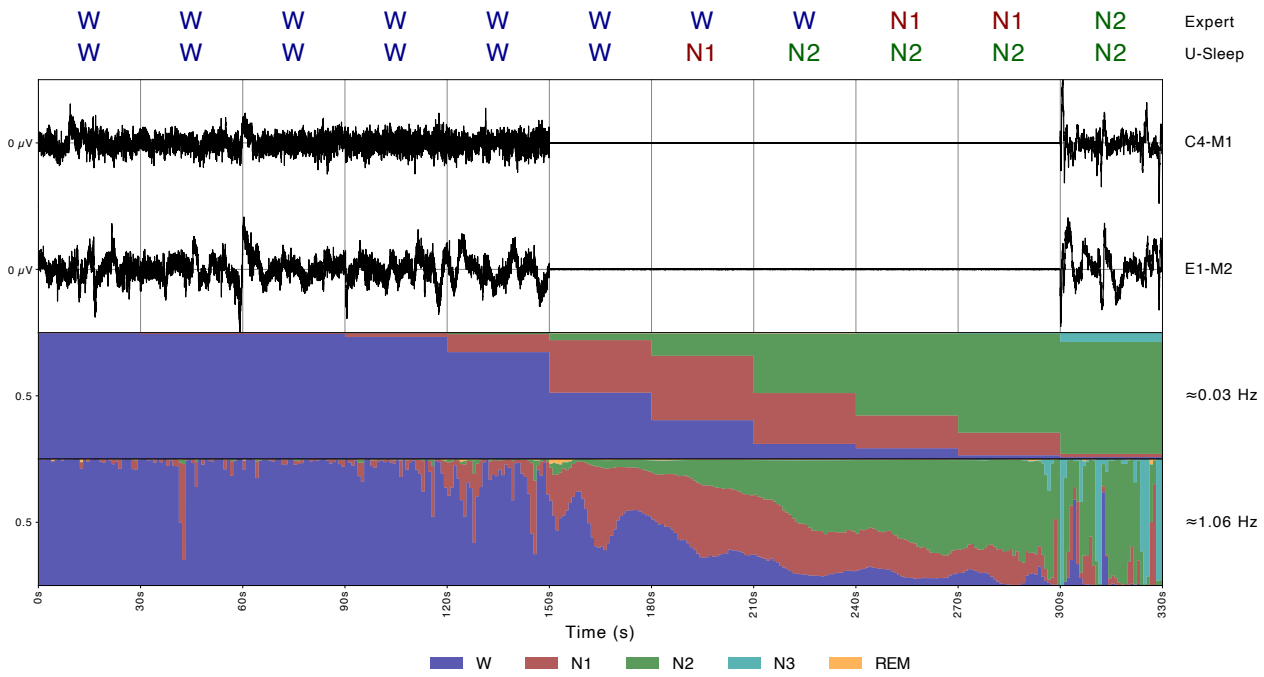


Figure F.4: Context Example 3. In this example, the central epoch and four proceeding epochs were removed from both input channels to the U-Sleep v2 model, which must score the epochs based only on distant neighbouring information. Notably, the model predicted a natural transition from the preceding Wake periods (for which information was available) to the preceding N2 epoch (for which information was available) via a transient N1 stage. However, the exact timing and length of the N1 transition were not correctly inferred. Note also that the confidence approximately linearly transitions from Wake to N1 to N2 within the region of missing data with intermediate N1 confidence.

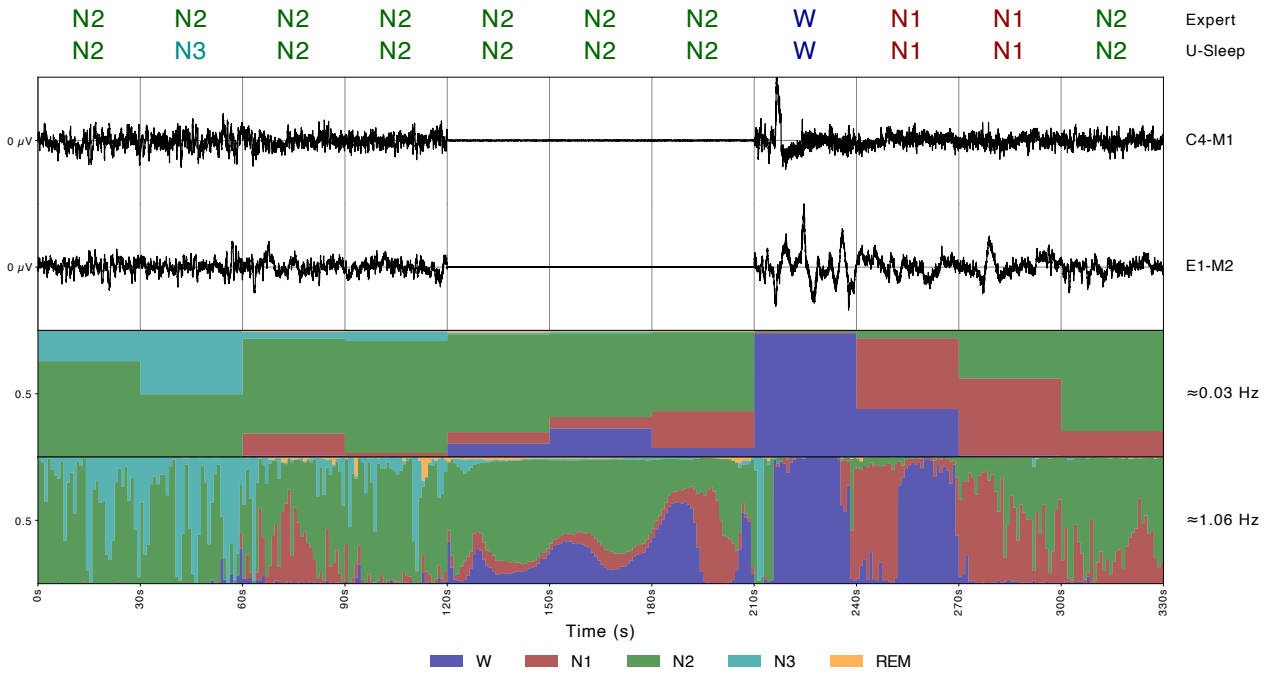


Figure F.5: Context Example 4. The three central-most epochs were replaced by noise in both input channels to the U-Sleep v2 model, which could correctly infer the most likely transition from N2 to Wake based on contextual information only. Note also that the model assigns increased confidence to stage N1 immediately before the Wake stage, indicating a possible preference for scoring the transition N2 → N1 → Wake over the direct N2 → Wake.

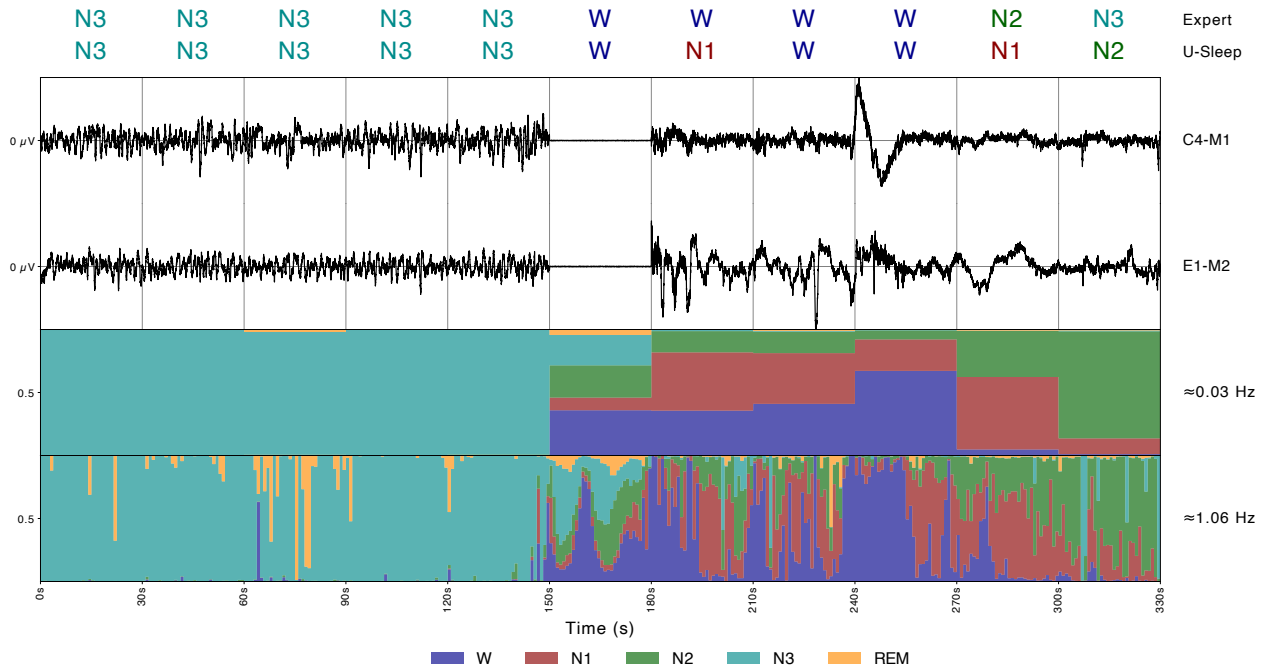


Figure F.6: Context Example 5. The central epoch was replaced by noise in both input channels to the U-Sleep v2 model. Despite the transition from a preceding period of stable N3 sleep directly to Wake occurring exactly at the epoch of missing information, the model correctly scored the transition based on the proceeding information, which indicates Wake stages.

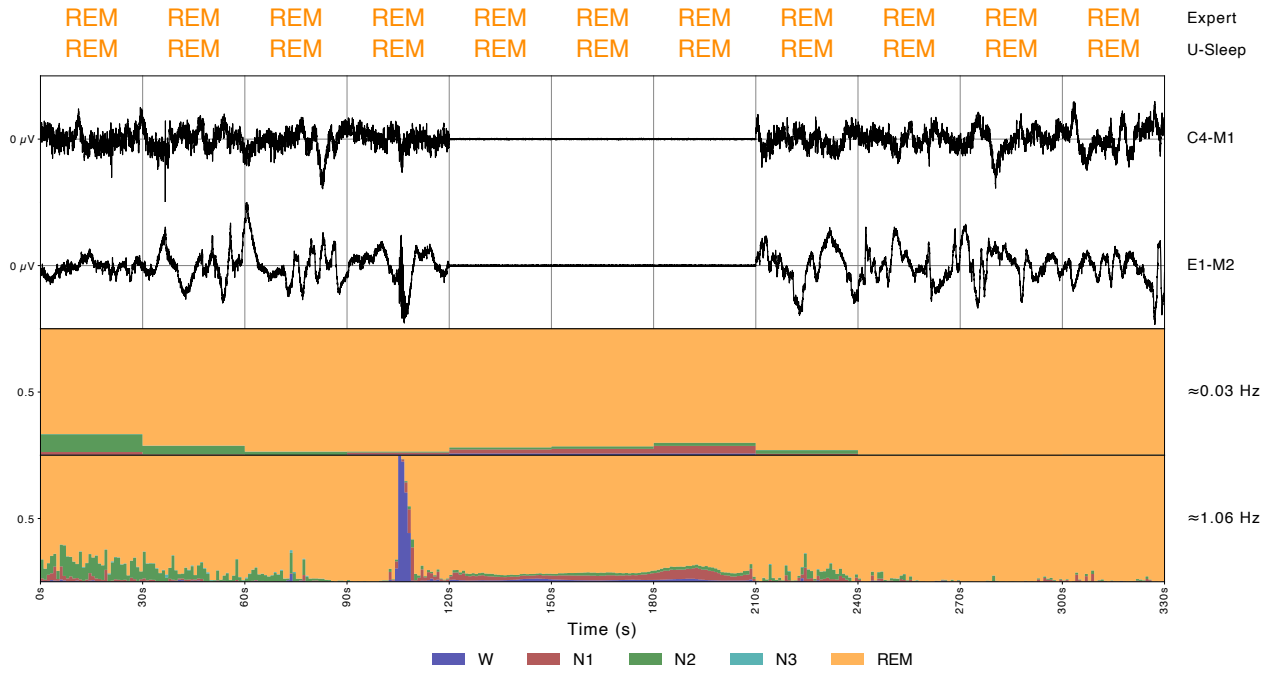
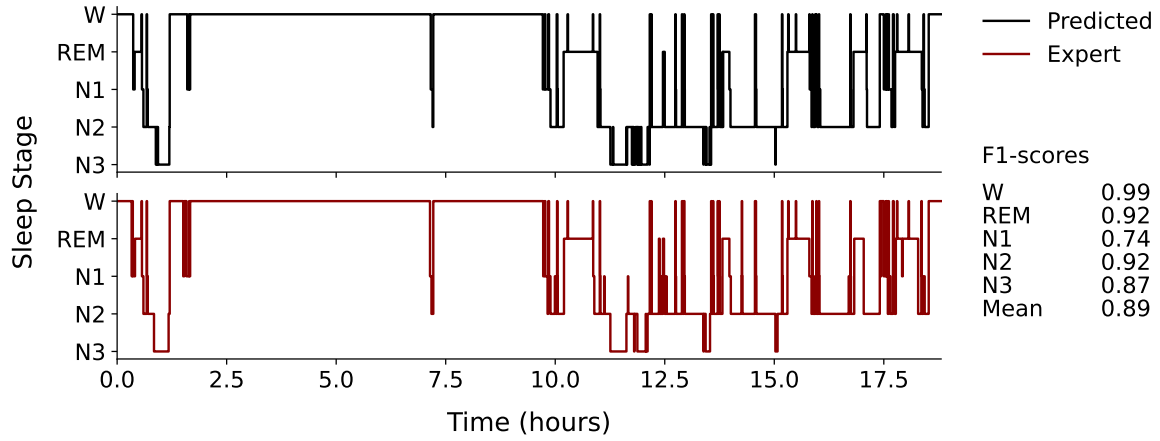
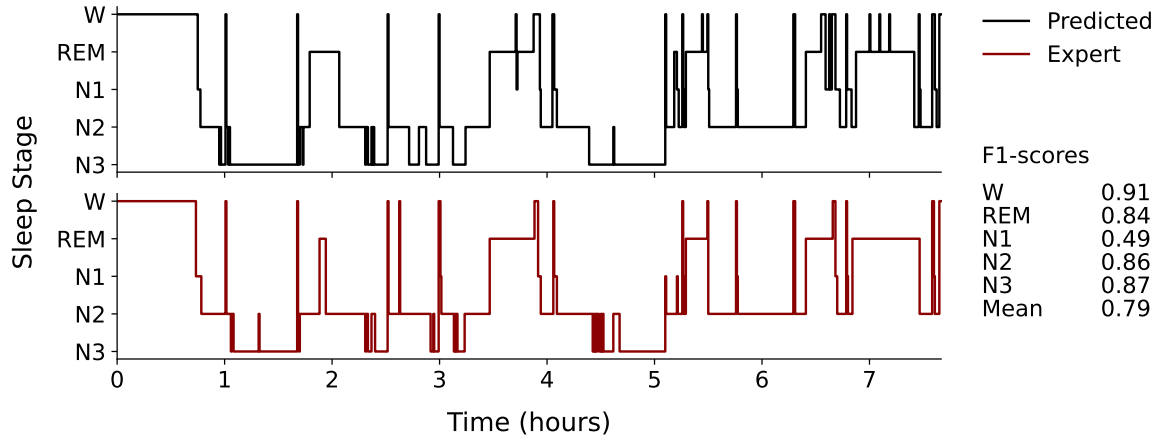


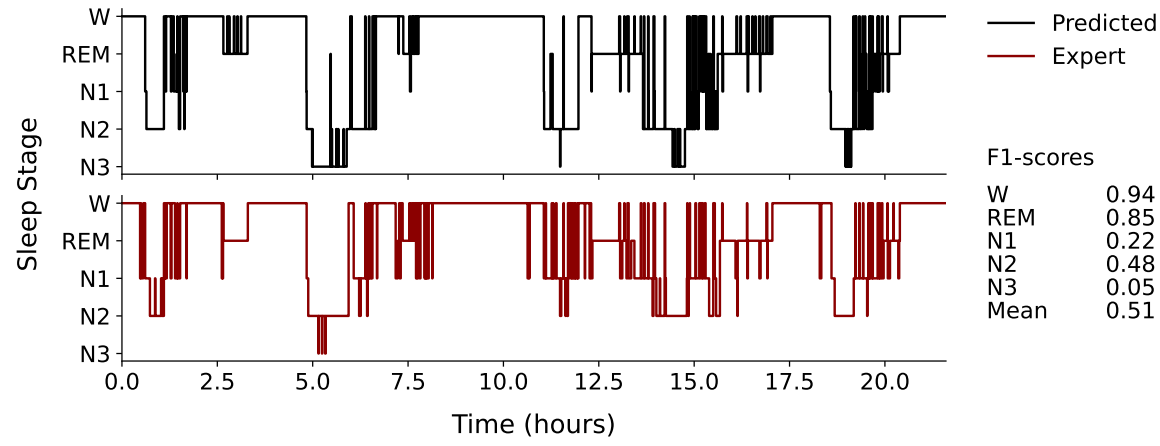
Figure F.7: Context Example 6. The central three epochs were replaced by noise in both input channels to the U-Sleep v2 model. In this example, a long stable period of REM sleep is easily scored by the model (with high REM stage confidence) despite the lack of 1.5 minutes of input data, because both pre- and proceeding information indicates a stable REM period.



(a) Hypnogram with highest observed macro F1-score (record `Sub104`).

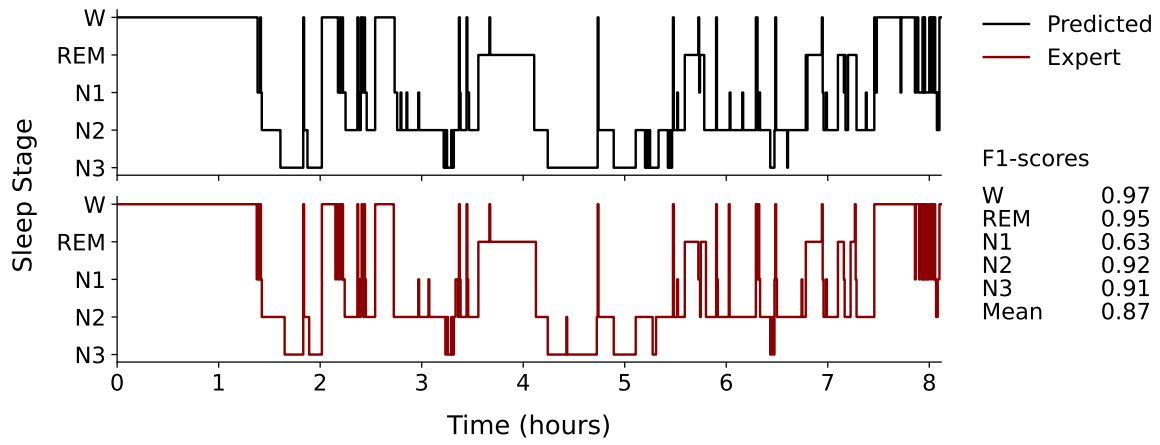


(b) Hypnogram with macro F1-score nearest dataset median (record `N0043`).

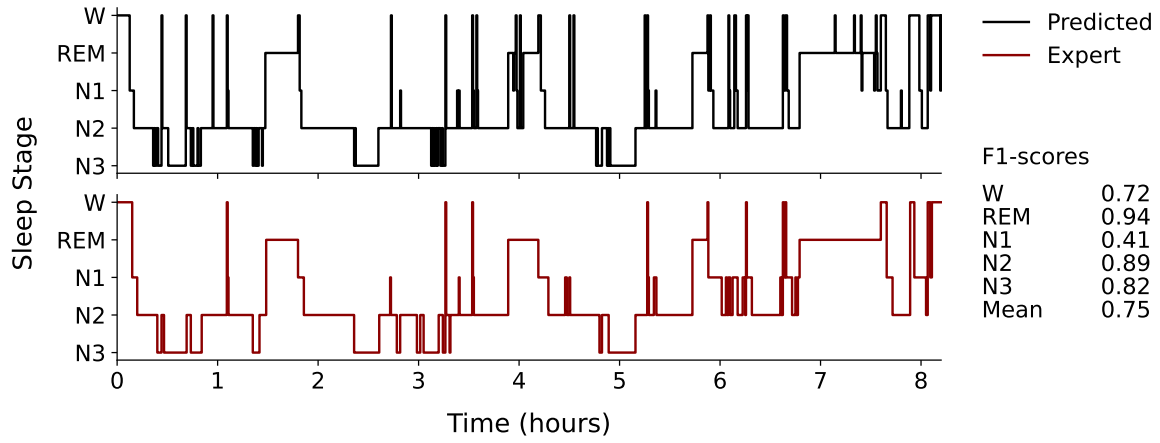


(c) Hypnogram with lowest observed macro F1-score (record `Sub125`).

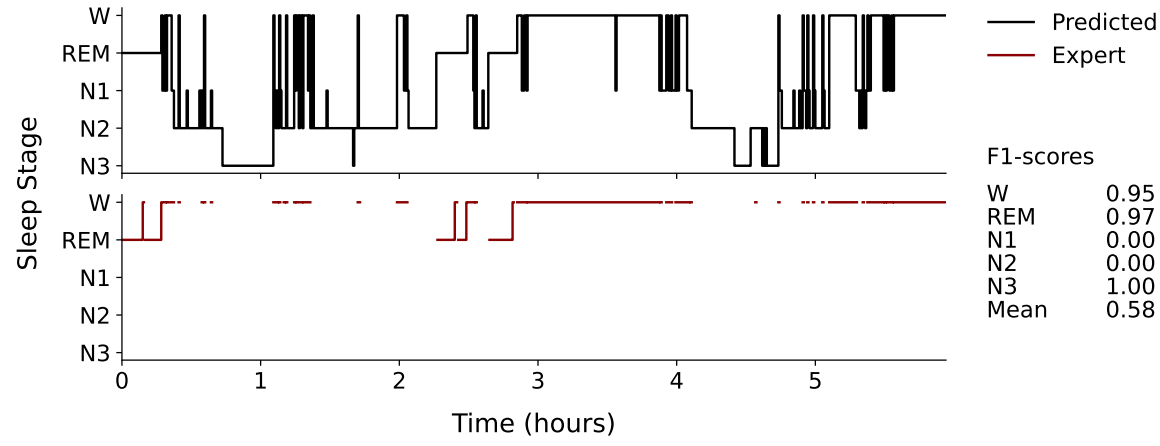
Figure F.8: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across dataset `DCSM-N`. Black hypnograms were predicted by U-Sleep, and red hypnograms are human expert annotations. F1 scores for each stage are shown to the right. Note that these unweighted per-subject F1 scores are noisy and may be misleading if the human annotator scores only a few instances of a given stage.



(a) Hypnogram with highest observed macro F1-score (record PLM117).

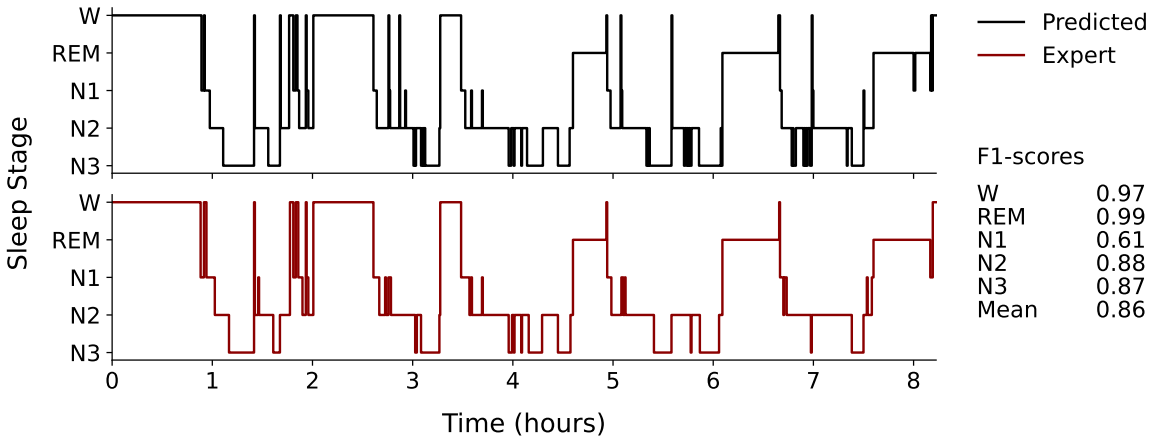


(b) Hypnogram with macro F1-score nearest dataset median (record PLM101).

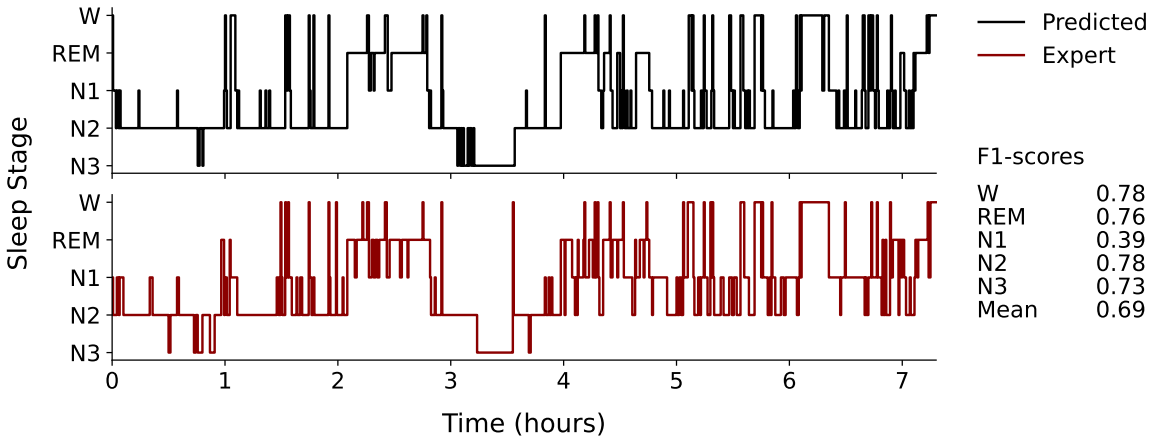


(c) Hypnogram with lowest observed macro F1-score (record PLM122). The missing areas of the expert's hypnogram indicate that a stage was not assigned, making the computed F1 scores highly variable with small changes in the numbers of correctly or wrongfully predicted stages for the classes with few ground truth instances.

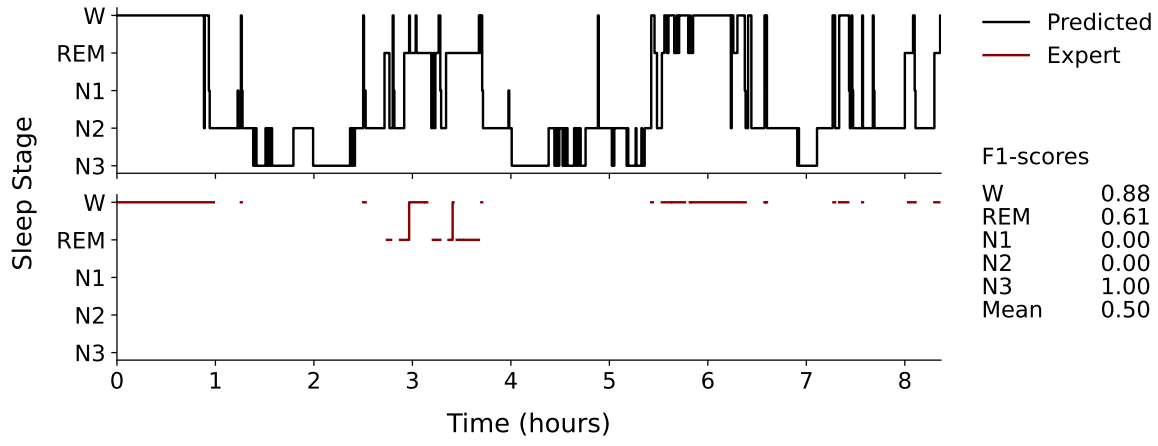
Figure F.9: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across dataset DCSM-PLM. Black hypnograms were predicted by U-Sleep, and red hypnograms are human expert annotations. F1 scores for each stage are shown to the right. Note that these unweighted per-subject F1 scores are noisy and may be misleading if the human annotator scores only a few instances of a given stage.



(a) Hypnogram with highest observed macro F1-score (record RBD097).

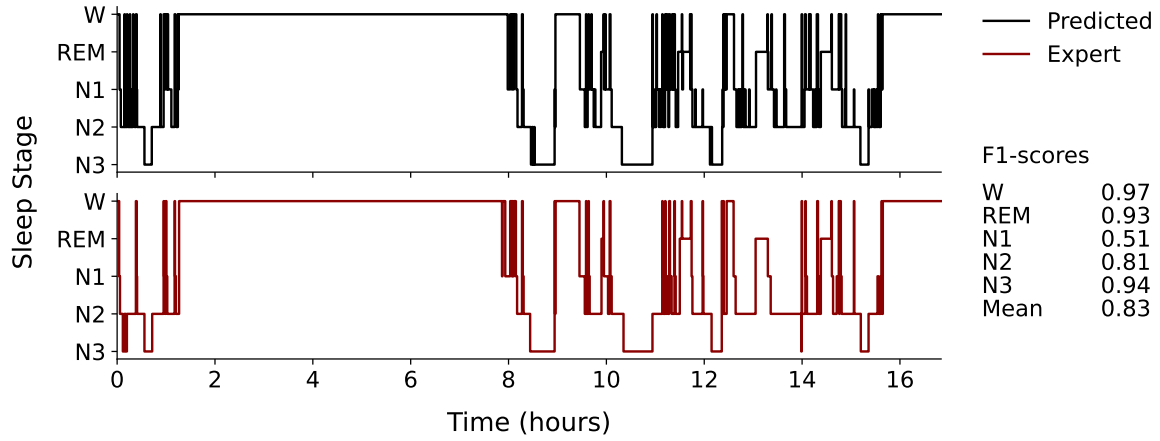


(b) Hypnogram with macro F1-score nearest dataset median (record RBD059).

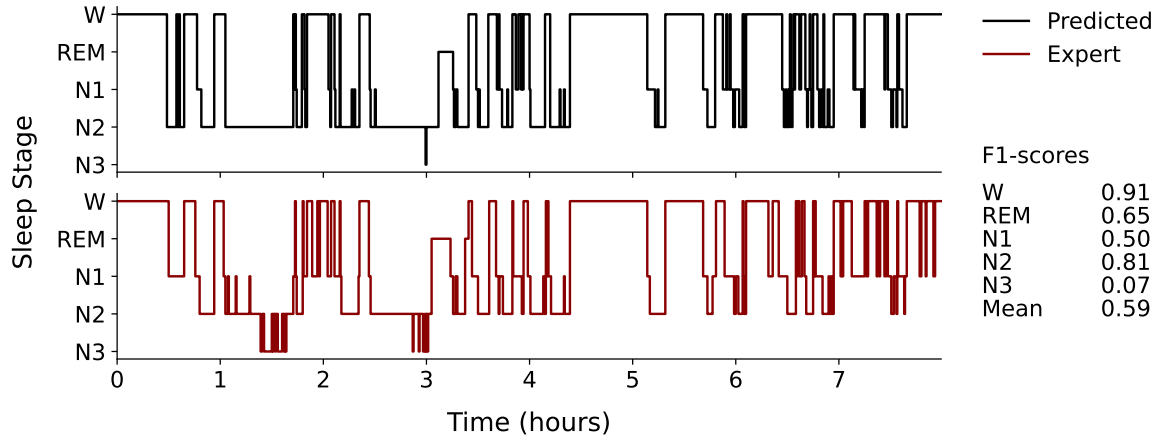


(c) Hypnogram with lowest observed macro F1-score (record RBD089). The missing areas of the expert's hypnogram indicate that a stage was not assigned, making the computed F1 scores highly variable with small changes in the numbers of correctly or wrongfully predicted stages for the classes with few ground truth instances.

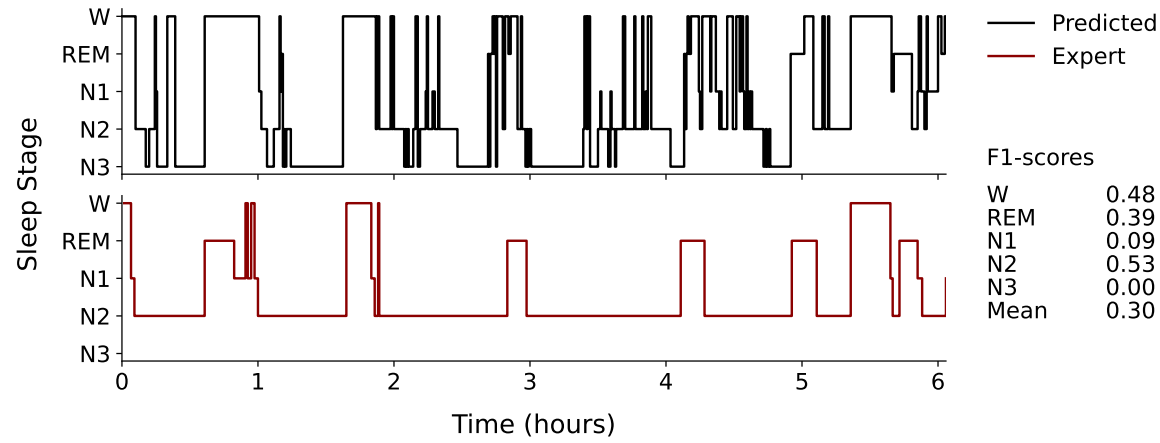
Figure F.10: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across dataset DCSM-RBD. Black hypnograms were predicted by U-Sleep, and red hypnograms are human expert annotations. F1 scores for each stage are shown to the right. Note that these unweighted per-subject F1 scores are noisy and may be misleading if the human annotator scores only a few instances of a given stage.



(a) Hypnogram with highest observed macro F1-score (record RBD-PD-050).



(b) Hypnogram with macro F1-score nearest dataset median (record RBD-PD094).



(c) Hypnogram with lowest observed macro F1-score (record RBD-PD-012).

Figure F.11: Highest, nearest median and lowest scoring (majority voted) hypnograms observed across dataset DCSM-RBD-PD. Black hypnograms were predicted by U-Sleep, and red hypnograms are human expert annotations. F1 scores for each stage are shown to the right. Note that these unweighted per-subject F1 scores are noisy and may be misleading if the human annotator scores only a few instances of a given stage.

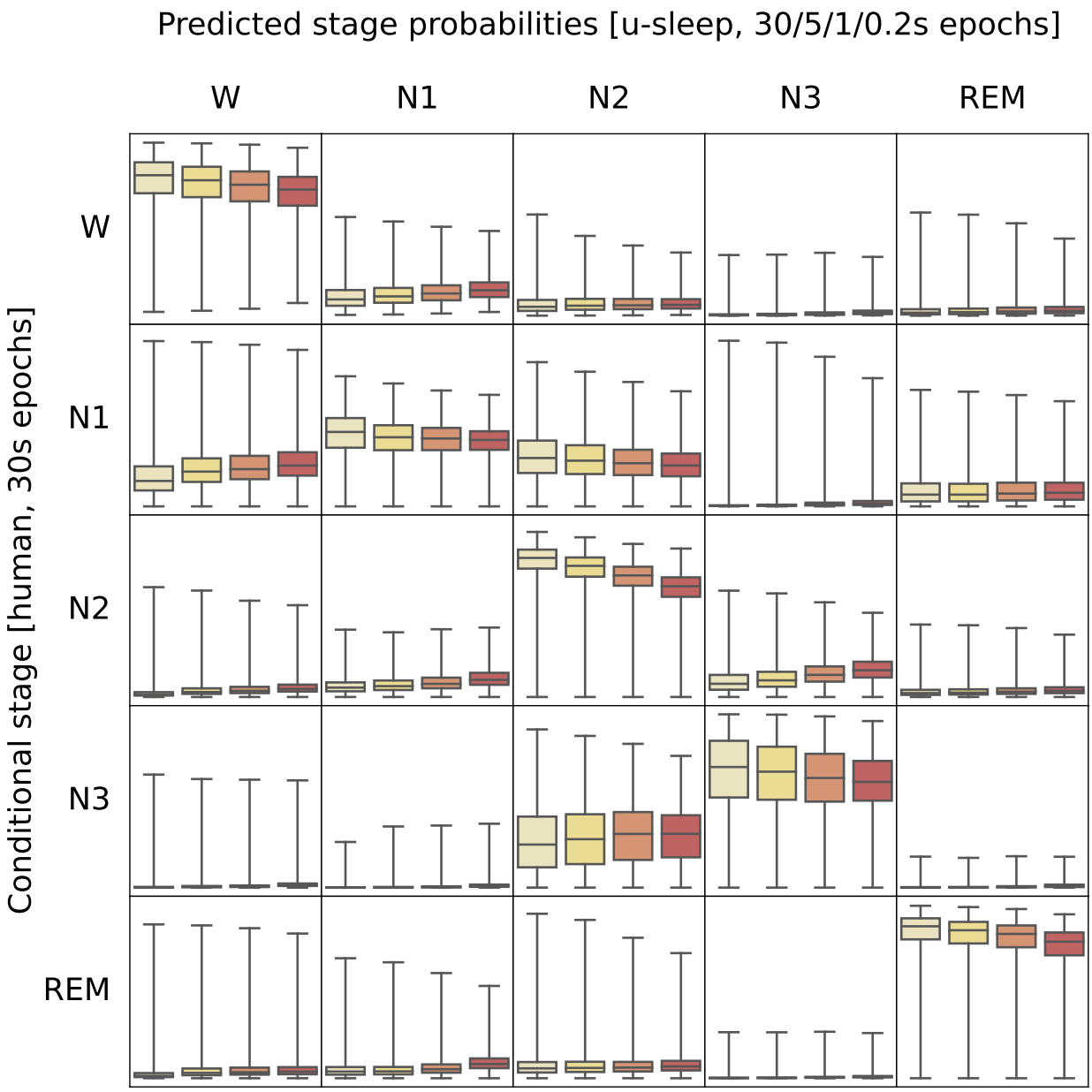
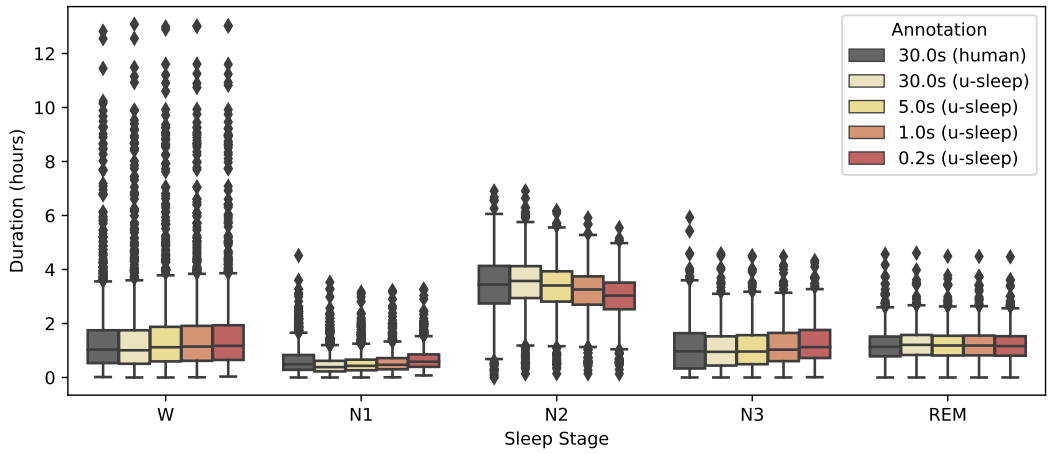
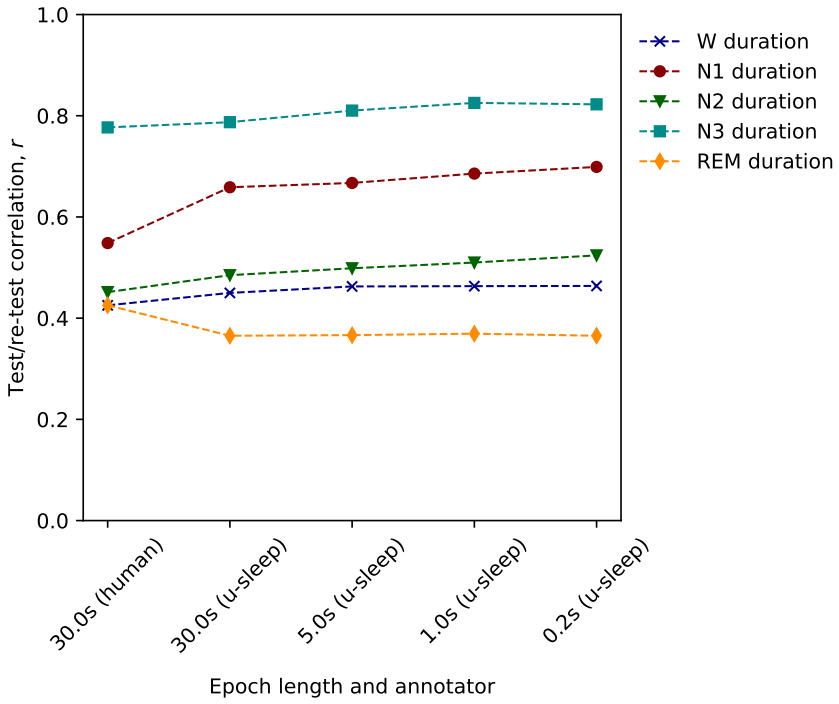


Figure F.12: Confusion matrix for U-Sleep v2 at different staging frequencies as applied to the entire set of $N = 2499$ PSGs across all validation- and test-set splits. Each row corresponds to the five sleep stages derived by human experts in 30-second intervals. The sum of confidence scores output by the model corresponding to each of the five sleep stages within the conditional segments is plotted in each of the five columns. Each boxplot within a single cell figure represents scores computed at different staging frequencies (from left to right: 1/30 Hz, 1/6 Hz, 1 Hz and 5 Hz). For instance, in the diagonal cells, boxplots show the sum of confidence scores at different frequencies that the model assigns to the human expert's derived stage within windows of 30 seconds. The plot in the second cell in the first row shows confidence that the model assigns to stage N1 across all windows scored as Wake by human experts. Boxplot show minimum, Q1, median, Q3 and maximum values (outliers not separately plotted for visual clarity).

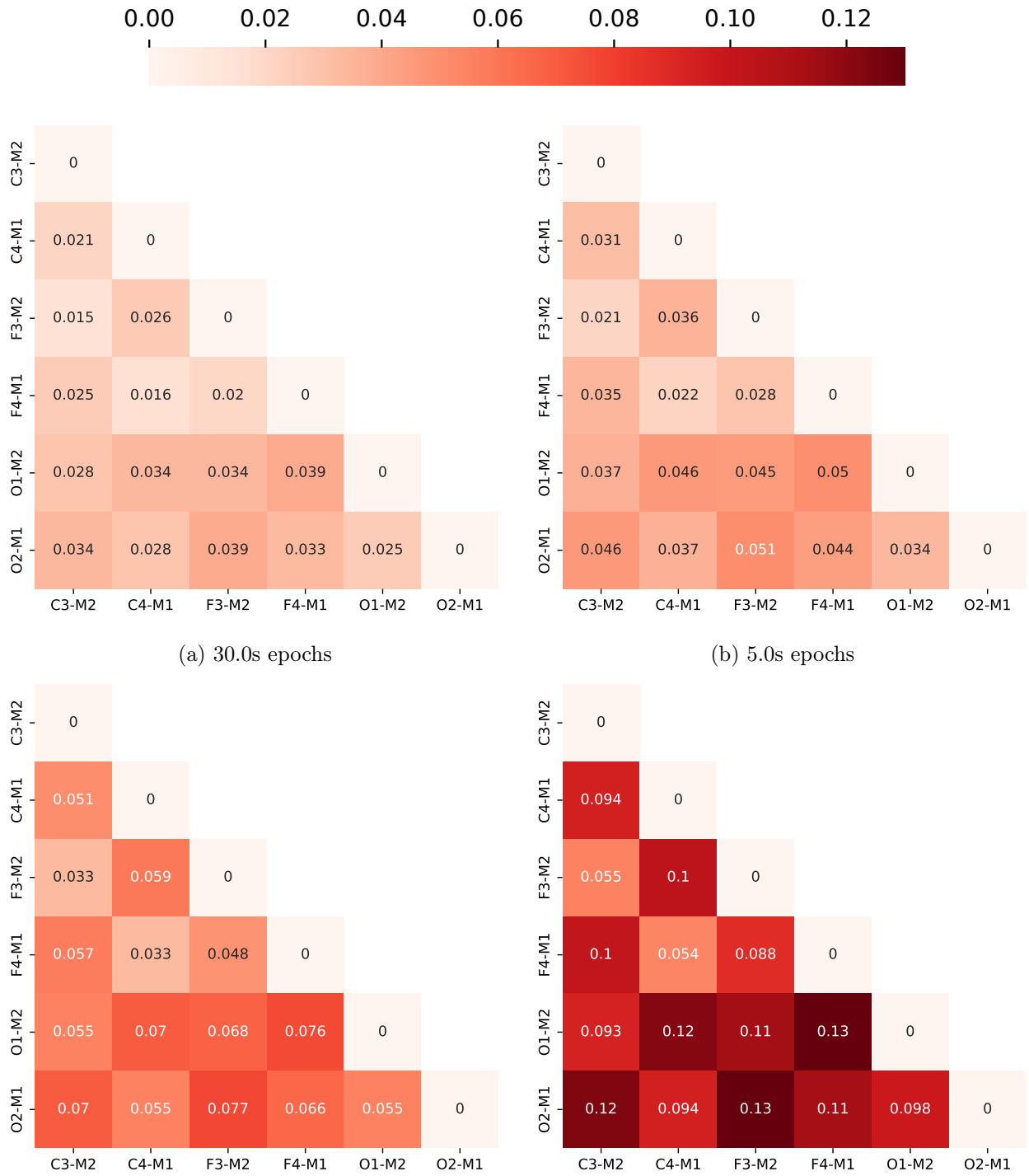


(a) Stage durations (hours)



(b) Test/re-test correlations, r.

Figure F.13: Visual representations of data in Table 11.3. Shows sleep stage durations as a function of staging frequency for human annotators and U-Sleep v2 and corresponding test/re-test Pearson's correlation r values for stage duration estimates made by different annotators.



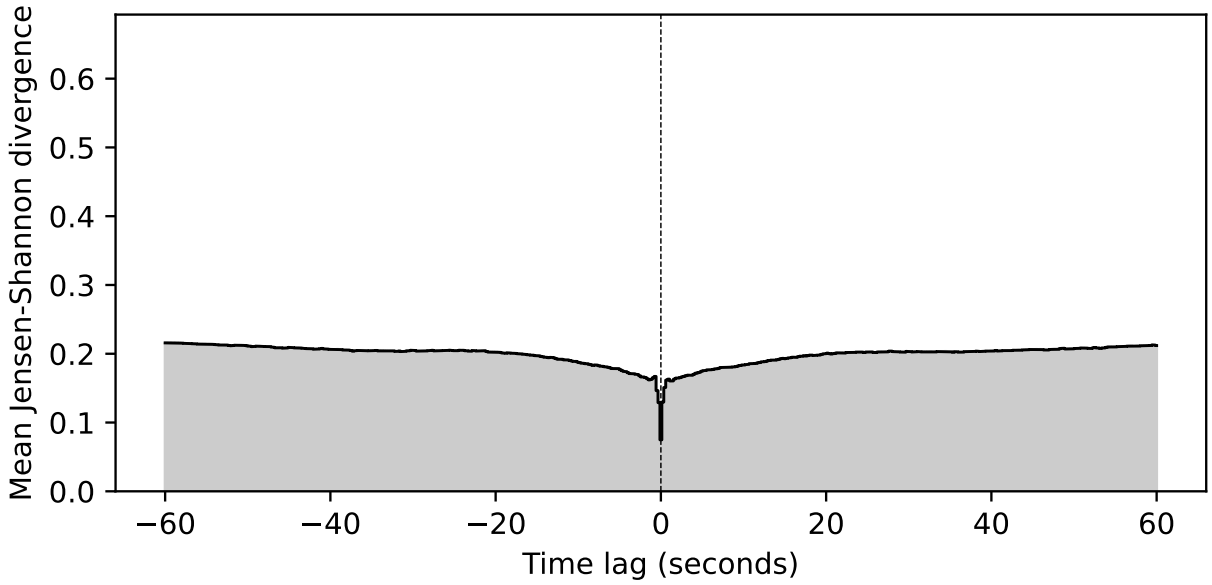
(a) 30.0s epochs

(b) 5.0s epochs

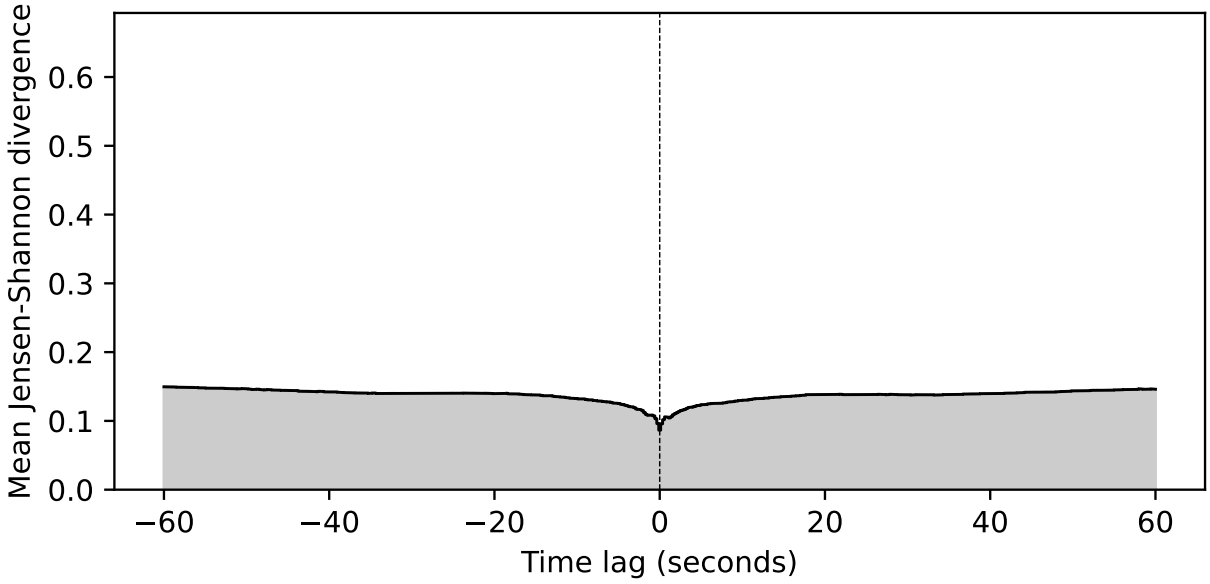
(c) 1.0s epochs

(d) 0.2s epochs

Figure F.14: Spatial sleep scoring divergence (dis-similarity) experiment using the U-Sleep v2 (EEG) model. Each number shows the mean epoch-wise Jensen-Shannon divergence (JSD) measure between the probability-like confidence score outputs of the model made in a pair of channels at different frequencies. Higher numbers (darker red) indicate that predictions were more diverging (i.e., less similar) on average. All $N = 853$ validation- and test-set PSGs with all channels available were used. Note that the colour bar is capped to the range $[0, 0.13]$ to highlight differences, although the mean JSD measure in this experiment, in principle, could take a maximum value of $\log 2 \approx 0.693$ if orthogonal predictions were made in all epochs in all PSGs.



(a) Example 2. Mean JSD cross-correlation experiment for one randomly selected PSG and pair of EEG channels (O1-M2 and F3-M2, with common E1-M2). The lowest JSD at lag 0 seconds indicate that the two sequences were most similar when not shifted relative to each other, i.e., there was no fixed lag effect



(b) Example 2. Mean JSD cross-correlation experiment for one randomly selected PSG and pair of EEG channels (O2-M1 and C3-M2, with common E1-M2). The lowest JSD at lag 0 seconds indicate that the two sequences were most similar when not shifted relative to each other, i.e., there was no fixed lag effect.

Figure F.15: Cross-correlation-like experiment measuring the mean Jensen-Shannon Divergence (see Methods) between two predicted sleep staging sequences at 5 Hz using the U-Sleep v2 model shifted by ±60 seconds relative to each other. The experiment showed that the JSD is minimized in nearly all cases at shift 0, i.e., no shift, indicating no fixed lag effect between the two predicted sequences. Two randomly selected examples are shown above.