# Automated Sleep Staging from Wearables and from Vision using Deep Learning




Jonathan Carter

Magdalen College

University of Oxford

Supervised by

Professor Lionel Tarassenko

A thesis submitted for the degree of

*Doctor of Philosophy*

Michaelmas 2024

# Acknowledgements

Firstly, I must thank my fantastic supervisor, Lionel. I hope that my numerous colloquialisms can be spared the red pen here. Thank you for always making time for me, despite a schedule which gets busier every week. Under your supervision, I've learnt a lot about how to conduct impactful research and a somewhat unexpected amount about English grammar–I'll use my Welsh upbringing as an excuse there. Above all, I am grateful for your trust and belief in me to solve problems in my own way, and for your clear vision and motivation of the bigger picture that has helped me to stay focused on what truly matters. I leave Oxford with a growing list of research questions and concrete ideas of how to address them. I hope we can continue to answer many of these together.

**Llandissilio (-2016).** In many ways, the world I now find myself in feels a million miles away from the tranquil part of the UK where I grew up. I am indebted to my parents, whose hard work and selflessness laid the foundations so that a boy from rural Wales could be the first in his family to become a Doctor, albeit not a particularly useful one for a medical emergency. Although I'm not able to get home nearly as much as I like to, I'm lucky that when I do, I'm often able to spend it with siblings (Oliver and Siobhan) and childhood friends (Tom M, Will, Tom W, Ewan, Ross, Fabian and many others) who all mean the world to me.

**Cambridge (2016-2020).** It seems dangerous to reference 'The Other Place', but I doubt I'd be *here* if I weren't lucky enough to have been *there*. Amongst the many exceptional academics I had the privilege of learning from, I'd particularly like to thank Prof. Tom Hynes for his teaching and guidance and Prof. Tony Purnell for his mentorship working with British Cycling in the summer of 2018. My only academic regret in four years at Cambridge was not doing my final-year project with Tony and the 'secret squirrel club'.

**Oxford (2020-2024).** Trading in one archaic collegiate institution for another, I'm very grateful to have been joined by old friends, including Izzy, Solomon, and Will McClune, and by new friends along the way. To the residents of The Palace, 122 Howard Street (Al+Rosie, Jake, Rob and Shreshth), I wish every year could be filled with as much laughter and delicious food as my last 12 months in Oxford.

My time in Oxford would not have been possible without the support of Oli Gibson, JJ Dixon, and Hugh Lloyd-Jukes, who, back in 2018, put their faith in

someone who didn't know what the number 14 was in Roman numerals just a few months earlier.[1] I'm also immensely grateful to so many former colleagues at Oxehealth (Bernardo, Bindia, David, Fran, João, Krzysztof, Matt, Ron, Simon and more) from whom I learnt so much during my time there. Studying alongside exceptionally talented colleagues at the AIMS CDT has been awesome. This is in no small part thanks to Wendy Poole, who is the beating heart of our CDT. I also owe a special shout-out to Cara for being the best host (even in absentia) on all three occasions I've been fortunate enough to visit North America during my DPhil. I can't wait for our next Big Day Out™.

**London (2024-)** The past four years in Oxford have been the best of my life but, whilst I am sad to have left, the upside of moving to London has been spending more time with wonderful friends (Ethan, Jo, John, Hattie, Lewis, Natalie, Ben HT, Ben J, Dani, Emma and many more) from Cambridge and beyond. It's been particularly exciting to be reunited with Will Halfpenny, my favourite person in the world to nerd out with.

Last, and opposite of least, Ella. Thank you for being my biggest believer, Minister of Special Treats™, and for your ruthless deletion of commas. Thank you for the time you made a five-hour round trip to rescue me from Oxford just after passing your driving test, and for the many roads (and roundabouts) we've navigated together since then. My life is better along every axis because of your unwavering support, brilliant sense of humour, and endless creativity. I love you.

[1] For completeness, it's XIV.

# Abstract

Quantitative sleep monitoring is essential for understanding its effects on health and well-being. Sleep medicine currently divides sleep into five stages, with the quantities and timings of these stages broadly linked to a variety of health conditions, ranging from dementia to depression. Unfortunately, existing methods for quantitative sleep monitoring are time-consuming, uncomfortable and expensive, requiring multiple contact sensors to be attached to the patient, and for a clinician to manually annotate the recorded data.

This thesis addresses these limitations by developing deep learning methods to automatically classify sleep stages from a reduced set of contact sensors (from wearables) or none at all using a near-infrared video camera (from vision). From wearables, we introduce a deep learning architecture for multi-modal time-series with several advantages over existing methods. Additionally, we introduce principled measures of model uncertainty and demonstrate how they could be used as part of the safe deployment of automated sleep monitoring in clinical practice. From vision, we show that multiple physiological parameters simultaneously derived from video data (cardiac, respiratory, and motion signals) can be combined to perform fully contact-free sleep stage classification (sleep staging). This is enabled by a novel deep transfer learning approach, which allows relationships between sleep stages and physiological signals learnt from contact sensors to be transferred to video with limited data.

The deep learning methods introduced in this thesis are trained and evaluated across multiple benchmark datasets spanning a wide range of demographics, including age, sex and ethnicity. Using an unprecedented amount of publicly available sleep data, we establish a new state-of-the-art for automated sleep staging accuracy across a range of less obtrusive sensing modalities. By taking a step towards affordable, non-invasive, and longitudinal sleep monitoring, these methods can improve our ability to diagnose and treat sleep disorders and help to further our understanding of the link between sleep and physical and mental health.

# Contents

**List of Figures** **vii**

**List of Tables** **ix**

**List of Abbreviations** **x**

**1 Introduction** **1**

1.1 Motivation . . . . . 1

1.2 Thesis Overview . . . . . 3

1.3 Contributions . . . . . 4

**2 Background and Related Work** **6**

2.1 Introduction . . . . . 7

2.2 Conventional sleep monitoring . . . . . 7

2.3 Deep learning . . . . . 14

2.4 Vision-based physiological measurement . . . . . 24

2.5 Automated sleep staging . . . . . 28

2.6 Conclusions . . . . . 34

**I Sleep Staging from Vision** **35**

**3 Video-based Sleep Staging using Transfer Learning** **36**

3.1 Introduction . . . . . 37

3.2 Related work . . . . . 40

3.3 Methods . . . . . 41

3.4 Results and Discussion . . . . . 49

3.5 Conclusions . . . . . 52

3.A Chapter Appendix . . . . . 53

**4 Video-based Sleep Staging from Waveforms** **57**
4.1 Introduction . . . 58
4.2 Background and Related Work . . . 60
4.3 Datasets and Preprocessing . . . 62
4.4 Methods . . . 64
4.5 Experimental Results . . . 69
4.6 Conclusions . . . 75
4.A Chapter Appendix . . . 76

**II Sleep Staging from Wearables** **84**

**5 Unified Multi-Modal Sleep Staging** **85**
5.1 Introduction . . . 86
5.2 Background and Motivation . . . 88
5.3 Model Architecture . . . 91
5.4 Experimental Set-up . . . 93
5.5 Results and Discussion . . . 97
5.6 Conclusions . . . 104
5.A Chapter Appendix . . . 105

**6 Automated Sleep Staging under Uncertainty** **109**
6.1 Introduction . . . 110
6.2 Background and Motivation . . . 111
6.3 Methods . . . 118
6.4 Experiments . . . 125
6.5 Conclusions . . . 137

**7 Conclusions** **141**
7.1 Review of Contributions . . . 141
7.2 Future Work . . . 143
7.3 Concluding Remarks . . . 146

**Appendices**

**A Statements of Authorship** **148**

**B Polysomnography Databases** **153**
B.1 Database descriptions . . . 153
B.2 Discussion . . . 155

**Bibliography** **156**

# List of Figures


2.1 An example hypnogram showing sleep stages over a night's sleep. . 9
2.2 An example of a patient set-up for a polysomnography study. . . . 11
2.3 Example presentation of sleep stages in the electroencephalogram. 12
2.4 Comparison of linear and convolutional weight matrices. . . . . 17
2.5 Variants of the Transformer neural network layer . . . . . 20
2.6 Visualisation of Transformer attention matrices. . . . . 21
2.7 Visualising the human pulse using Eulerian Video Magnification . . 25

3.1 Sleep staging from vision using transfer learning. . . . . 39
3.2 Room and camera layouts used for data acquisition. . . . . 42
3.3 Extracting 1D activity signals from video using optical flow. . . . . 42
3.4 Coverage of video-derived vital-sign estimates. . . . . 45
3.5 Example vital signs extracted from video recorded during sleep. . . 46
3.6 Example vital signs measured from contact sensors during sleep. . . 47
3.7 Vital-sign representation learning from contact sensors. . . . . 48
3.8 Confusion matrix showing video-based sleep staging performance. . 50
3.9 Scatter plot of Cohen's $\kappa$ values against age and Fitzpatrick skin type. 50
3.10 Hypnograms corresponding to the median Cohen's $\kappa$ statistic. . . . 51
3.11 Hypnograms corresponding to the lowest Cohen's kappa statistic . . 55
3.12 Hypnograms corresponding to the lowest Cohen's kappa statistic . . 56

4.1 Sleep staging from near-infrared video signals using SleepVST. . . . 60
4.2 Example video-derived cardio-respiratory waveforms. . . . . 63
4.3 SleepVST architecture. . . . . 64
4.4 Waveform encoder design. . . . . 65
4.5 Near-infrared video frame processing. . . . . 67
4.6 Confusion matrix between SleepVST and expert labels. . . . . 71
4.7 Example generated hypnograms from the OSV dataset. . . . . 71
4.8 Model performance against age, sex and Fitzpatrick skin type . . . 72
4.9 Example four-class hypnograms during fragmented sleep. . . . . 74
4.10 Example processing of near-infrared video frames. . . . . 77
4.11 Performance against age across the SHHS, MESA and OSV datasets. 78

4.12 SleepVST confusion matrices for contact sensor datasets. . . . . 78
4.13 Additional example four-class sleep hypnograms from the OSV dataset. 80
4.14 Classifying longer input sequences using SleepVST . . . . . 82
4.15 Extracting features from longer input sequences using SleepVST. . . 82

5.1 Overview of wav2sleep . . . . . 87
5.2 wav2sleep model architecture . . . . . 91
5.3 Stochastic masking during wav2sleep training . . . . . 97
5.4 Confusion matrices for varying $\mathcal{S}_{\text{Test}}$ on the Census test set. . . . . 100
5.5 Performance against age for varying $\mathcal{S}_{\text{Test}}$ on the Census test set. . . 100
5.6 Performance grouped by ECG signal quality. . . . . 101
5.7 Model performance stratified by apnea severity and arrhythmia . . 102
5.8 Example hypnograms for a subject with cardiac arrhythmia . . . . 102
5.9 Effect of stochastic masking on model performance . . . . . 103

6.1 A statistical observation model for automated sleep staging. . . . . 112
6.2 Example of long-range inter-expert disagreement in sleep staging. . 115
6.3 Sleep staging under uncertainty using wearable signals. . . . . 115
6.4 SleepDecoder architecture. . . . . 119
6.5 Visualisation of attention masks used in the SleepDecoder model. . 120
6.6 Model calibration curves on the SHHS test set. . . . . 126
6.7 Performance-confidence frontiers for the MROS test set. . . . . 127
6.8 Confusion matrices for confident and unconfident sleep epochs. . . . 127
6.9 Comparison of confidence-accuracy trade-offs for uncalibrated and calibrated models. . . . . 129
6.10 Example night with low uncertainty from the MROS test set . . . . 130
6.11 Example night with high uncertainty from the MROS test set . . . 131
6.12 SleepDecoder models long-range uncertainty similarly to human experts. 132
6.13 Estimated sleep metrics over the DOD datasets. . . . . 132
6.14 Estimating sleep metric uncertainty over the DOD datasets. . . . . 133
6.15 Example ECG augmentations used to form the Dreem Augmented dataset . . . . . 135
6.16 Comparison of confidence-accuracy trade-offs for uncalibrated and calibrated models. . . . . 135

# List of Tables


2.1 Polysomnography equipment list. . . . . 10
2.2 Polysomnogram characteristics of specific sleep stages. . . . . 11
2.3 Definitions of common sleep metrics. . . . . 12
2.4 A rule of thumb for assessing Cohen's $\kappa$ . . . . . 29

3.1 Sleep study population demographics. . . . . 41
3.2 Motion features derived from optical flow. . . . . 43
3.3 Coverage statistics for video-derived vital signs. . . . . 45
3.4 Comparison with prior video-based sleep staging work. . . . . 49
3.5 Performance across different sleep stage stratifications. . . . . 51
3.6 Performance improvements using motion and deep vital-sign features. 52
3.7 Ablation results showing the effect of SQI filtering. . . . . 52

4.1 Comparison of methods using contact sensors. . . . . 70
4.2 Comparison of four-class sleep staging methods using video cameras. 70
4.3 Performance for different model transfer strategies. . . . . 73
4.4 Ablation study of motion feature parameters. . . . . 73
4.5 Model performance for different cardio-respiratory inputs. . . . . 74
4.6 Model performance for different classification strategies . . . . . 79
4.7 Ablation of motion feature parameters $T, \delta$ and $\Delta$. . . . . 79
4.8 SleepVST transformer encoder parameters. . . . . 81

5.1 Overview of the datasets used. . . . . 94
5.2 wav2sleep model hyper-parameters. . . . . 96
5.3 Performance ($\kappa_T$) for $\mathcal{S}_{\text{Test}} = \text{PPG}$. . . . . 98
5.4 Performance ($\kappa_T$) for $\mathcal{S}_{\text{Test}} = \text{ECG}$. . . . . 99
5.5 Comparison of wav2sleep with prior work . . . . . 104
5.6 Performance for different test-time modalities. . . . . 105
5.7 Validation metrics for different epoch mixer designs. . . . . 107
5.8 Validation metrics for different sequence mixer designs. . . . . 107

6.1 Performance ($\kappa_T$) of different SleepDecoder variants. . . . . 125
6.2 Model-expert agreement statistics and expert-expert agreement statistics for the Dreem datasets. . . . . 136

# List of Abbreviations

**AASM** . . . . . American Academy of Sleep Medicine

**AHI** . . . . . Apnea-Hypopnea Index

**BR/RR** . . . . Breathing/respiratory rate (used interchangeably).

**CNN** . . . . . Convolutional neural network.

**ECG** . . . . . Electrocardiography.

**EEG** . . . . . Electroencephalography.

**EMG** . . . . . Electromyography.

**EOG** . . . . . Electrooculography.

**FDA** . . . . . U.S. Food and Drug Administration

**HR** . . . . . Heart rate

**MESA** . . . . . Multi-Ethnic Study of Atherosclerosis

**MHRA** . . . . U.K. Medicines and Healthcare products Regulatory Agency

**MLP** . . . . . Multi-layer perceptron.

**NIR** . . . . . Near-infrared.

**NSRR** . . . . . National Sleep Research Resource

**OSV** . . . . . Oxford Sleep Volunteers dataset

**PPG** . . . . . Photoplethysmography.

**PSG** . . . . . Polysomnography.

**REM** . . . . . Rapid Eye Movement

**SHHS** . . . . . Sleep Heart Health Study

**SQI** . . . . . Signal Quality Index.

**vPSG** . . . . . Video polysomnography.

# 1 Introduction

## 1.1 Motivation

Sleep, or sleep-like behaviour, is observed in virtually every animal species [190]. Experts recommend that human adults should aim to get between 7 and 9 hours of sleep each night [1]. However, data from the US indicates that almost 40% of adults do not get enough sleep [39], and that 50 to 70 million Americans suffer from a sleep disorder [142]. Poor sleep is associated with an alarmingly wide range of health conditions, and insufficient and excessive sleep have both been shown to be significant predictors of all-cause mortality [36]. Accurate, quantitative sleep monitoring is essential to improve our understanding of the relationship between sleep and physical and mental health, and to address these widespread issues through the development of sleep disorder treatments.

The 'gold standard' method for quantitatively assessing sleep is polysomnography, a test which requires a patient to tolerate a large number of uncomfortable contact sensors worn during sleep, the main signal recorded being the electrical activity of the brain at one or more sites on the scalp, the electroencephalogram (EEG). After an overnight recording, the recorded signals are manually reviewed by a sleep physiologist, a task which can take multiple hours to complete. A key component of this task is annotating sleep stages, which involves assigning each

30-second window of recorded data to one of five discrete states: Wake, N1, N2, N3 and Rapid Eye Movement (REM) sleep; where N1 to N3 account for varying degrees of non-REM sleep from near-wakefulness (N1) to deep sleep (N3). Metrics derived from these sleep stage annotations, such as the duration spent in REM sleep, can be used to derive biomarkers associated with conditions ranging from schizophrenia [120] to dementia [249].

Polysomnography is a time-consuming and costly clinical procedure which is uncomfortable for patients. This thesis aims to address these shortcomings by advancing the performance of automated sleep stage classification (sleep staging) from a reduced set of contact sensors. Doing so could have benefits for important clinical applications such as:

- Accelerating the discovery of treatments for sleep disorders such as insomnia, by enabling faster and cheaper clinical trials.
- Longitudinal monitoring of *known* sleep biomarkers to detect the presence of underlying conditions. For example, a decrease in deep (N3) sleep is associated with Alzheimer's disease [249].
- The discovery of *novel* sleep biomarkers, using much larger study populations than currently feasible using polysomnography.

To achieve this, we aim to use data-driven, deep learning-based approaches that are general and modality-agnostic, enabling them to be applied to many different modalities used within healthcare settings with minimal changes. We broadly focus on automated sleep staging from two categories of less obtrusive sensing modalities: wearables and vision. Video data is an important modality in sleep medicine, and is commonly used for applications such as the diagnosis of sleep movement disorders [206]. Accurate sleep staging solely from video data could enable it to be used as the sole input modality for performing a comprehensive sleep assessment with minimal equipment set-up or disruption to the patient. Meanwhile, the ubiquity and low cost of wearable sensors means they could be used at scale for the longitudinal

monitoring of sleep in large study populations, to understand how sleep affects human health and wellbeing over the course of days, weeks, months and years.

## 1.2 Thesis Overview

This thesis is presented in an integrated format, comprising a mix of traditional thesis chapters and paper-based chapters, which have been stylistically integrated into the thesis. **Chapter 2** first provides an overview of conventional sleep monitoring and reviews the existing literature on automated sleep staging, vision-based physiological measurement, and relevant deep learning methods. **Chapters 3 to 6** are then divided into two parts, described below, where paper-based chapters are denoted by an asterisk (*). Finally, **Chapter 7** concludes this thesis with reflections and suggestions for future work.

**Part I - Chapters 3* and 4* - Sleep Staging from Vision**

This part of the thesis builds upon advances in the field of vision-based physiological measurement to perform automated sleep stage classification entirely from near-infrared video. **Chapter 3** introduces a deep transfer learning approach to video-based sleep staging and demonstrates how multiple physiological parameters measured from video: the heart rate, breathing rate, and movement; can simultaneously be used to classify sleep stages. **Chapter 4** extends these ideas by using cardiac and respiratory waveforms as model inputs, rather than derived heart and breathing rates, and by utilising a Transformer neural network architecture [228].

**Part II - Chapters 5* and 6 - Sleep Staging from Wearables**

In the second part of this thesis, we return to contact sensor data to address two important questions:

Q1. How can we further advance the performance of sleep staging algorithms from non-EEG sensors, such as wearables or video?

Q2. Beyond simply improving raw accuracy, how can we use uncertainty quantification to enable real-world applications of non-EEG sleep staging?

To address Question 1, **Chapter 5** introduces a novel deep learning architecture that can operate on a variable set of inputs during training and inference. This allows a model to be jointly trained across multiple sleep datasets, for which the availability of the inputs can vary. This increases the scale and variety of available data, which is crucial to the success of deep learning-based approaches [67]. To address Question 2, **Chapter 6** explores how uncertainty quantification could enable applications of non-EEG sleep staging without requiring continuous expert-level accuracy as measured by conventional metrics.

## 1.3 Contributions

The work described herein has led to the publication of three conference papers, the filing of two patents, and regulatory clearance of a world-first medical device for contact-free sleep monitoring by the US Food and Drug Administration (FDA).

The methods introduced in **Part I** advance the state-of-the-art in video-based sleep staging, outperforming prior methods that have previously [153] and concurrently [100, 226, 239] addressed this problem. By combining multiple physiological signals measured from video, this part of the thesis shows that sleep staging can be performed with near-expert-level accuracy entirely contact-free from a near-infrared video camera.

Returning to contact sensor data in **Part II**, we simultaneously advance the state-of-the-art in sleep staging accuracy across a range of input modalities such as the ECG, PPG, and respiratory signals, which can be acquired using modern wearable devices. These advances are enabled by the development of a single, unified model that can operate on an arbitrary set of inputs during training and inference, which is introduced in Chapter 5. Beyond the problem of sleep staging addressed in this thesis, this model has broader potential applications in the healthcare domain, where heterogeneous datasets are especially common, and could benefit emerging areas of research such as unsupervised representation learning from physiological signals [2, 215, 258]. The training code and model weights from this project

have been open-sourced, enabling the wider research community to apply these algorithms to their data and to easily build on these advances.

Finally, Chapter 6 discusses how uncertainty quantification could be used to overcome inherent limitations of automated sleep staging from modalities such as wearables or video. This chapter addresses a flawed modelling assumption used in almost all prior sleep staging work, enabling the derivation of principled measures of uncertainty. The resulting model can produce reliable measures of confidence over its outputs and we demonstrate how it could be used to enable applications of wearable sleep staging such as sleep disorder screening or evaluating treatment efficacy.

Alongside the development of this thesis, the algorithms introduced in Chapters 3 and 4 were translated into Software as a Medical Device (SaMD), 'Oxevision Sleep', which obtained regulatory clearance from the US Food & Drugs Administration (FDA) for contact-free sleep monitoring in early 2024 [219]. At the time of writing, the algorithms introduced in Chapters 5 and 6 are being evaluated in ongoing clinical studies in Oxford, using signals recorded from various wearable devices.

# 2

# Background and Related Work

## Contents


**2.1 Introduction . . . . . 7**
**2.2 Conventional sleep monitoring . . . . . 7**
2.2.1 Physiological basis . . . . . 7
2.2.2 Polysomnography . . . . . 9
2.2.3 Limitations . . . . . 13
**2.3 Deep learning . . . . . 14**
2.3.1 Multi-layer perceptrons . . . . . 14
2.3.2 Supervised learning . . . . . 15
2.3.3 Convolutional neural networks . . . . . 16
2.3.4 Transformers . . . . . 18
2.3.5 Transfer learning . . . . . 23
**2.4 Vision-based physiological measurement . . . . . 24**
2.4.1 Cardiac monitoring . . . . . 24
2.4.2 Respiratory monitoring . . . . . 26
2.4.3 Real-world use and validation . . . . . 26
2.4.4 Video actigraphy . . . . . 27
2.4.5 Future additions . . . . . 27
**2.5 Automated sleep staging . . . . . 28**
2.5.1 PSG-based sleep staging . . . . . 30
2.5.2 Wearable sleep staging . . . . . 30
2.5.3 Non-contact sleep staging . . . . . 31
2.5.4 Foundational assumptions . . . . . 33
**2.6 Conclusions . . . . . 34**

## 2.1 Introduction

This chapter introduces relevant background theory and describes important prior work. Section 2.2 first describes relevant aspects of sleep physiology and reviews conventional methods for sleep monitoring, highlighting their limitations. Section 2.3 provides an overview of the deep learning methods used in the later chapters of this thesis. Section 2.4 then reviews the field of vision-based physiological monitoring, which underpins the work of Chapters 3 and 4. Finally, Section 2.5 introduces prior work on the problem of automated sleep staging.

## 2.2 Conventional sleep monitoring

### 2.2.1 Physiological basis

Sleep and wakefulness are not binary states, a fact which has been known since at least the mid-19th century, when Kohlschütter first studied the effect of acoustic stimuli on sleep depth [107]. However, most of our understanding of sleep has occurred since the 1920s after the invention of the EEG, which measures the electrical activity of the brain at one or more sites on the scalp. This quickly enabled overnight monitoring of brain activity, helping to further our understanding of the role of specific brain sites in the regulation of sleep [189].

The next major advancement in our understanding of sleep occurred shortly after the invention of the electrooculogram (EOG) in 1951 [138]. This quickly led Aserinksy and Kleitman [12] to discover the existence of a distinct phase of sleep which we now refer to as Rapid Eye Movement (REM) sleep. During this phase of sleep, the eyeballs move rapidly underneath the (closed) eyelids, giving rise to the name of this phase. The brain's electrical activity (as measured by the EEG) is similar to that during wakefulness, leading to REM sleep being described as 'paradoxical' [99]. Changes in brain signalling induce muscle atonia—temporary paralysis which prevents us from acting out our dreams, which are now known to be more vivid and more intense during REM sleep [211].

**Sleep stages.** In 1968, Rechtschaffen and Kales [176] formalised these observations into the 'R&K manual', a set of rules to divide sleep into discrete states based on the visual analysis of plots of physiological measurements from sensors including the EEG and EOG. These rules divided sleep into 6 discrete states: Wake, S1, S2, S3, S4 and REM sleep; where S1 to S4 account for varying degrees of non-REM sleep from near-wakefulness (S1) to deep 'slow-wave' sleep (S4).

Today, those rules have been superseded by the American Academy of Sleep Medicine (AASM) guidelines [94], which were designed to simplify those of R&K. Given ambiguity in distinguishing S3 and S4, the AASM rules merged S3 and S4 into a single state, N3, with N1 and N2 replacing the original S1 and S2 states from the R&K guidelines.

**Sleep cycles.** Following the AASM stratification, in the first 15-30 minutes of normal sleep, we gradually progress from N1 to N2 to N3 sleep. As we do so, muscle activity decreases [197] and the cardiac and respiratory rates gently decrease and reduce in variability [32, 73]. From N3 sleep, we then transition to REM sleep. This cycle through the sleep stages typically takes around 90 minutes for an adult, and over the course of a night's sleep we typically undergo four to six sleep cycles [161].

In practice, normal sleep usually deviates slightly from this idealised behaviour. Sensory stimulation such as environmental noise may cause transitions back to lighter stages of sleep or wakefulness. Stimulants such as alcohol [214], caffeine [22] or medications [157] can also cause sleep disruption.

**Sleep hypnograms.** The transition between sleep stages over the course of a night's sleep can be presented graphically on a plot known as a sleep hypnogram, an example of which is shown in Figure 2.1. Here we can observe five REM cycles over the course of the night. We can additional observe occasional, brief awakenings (arousals) throughout the night, which are typically not remembered after waking up. Arousals are a normal part of sleep which occur in response to internal and environmental stimuli, they "ensure the reversibility of sleep, without which it would be identical to [a] coma" [75].

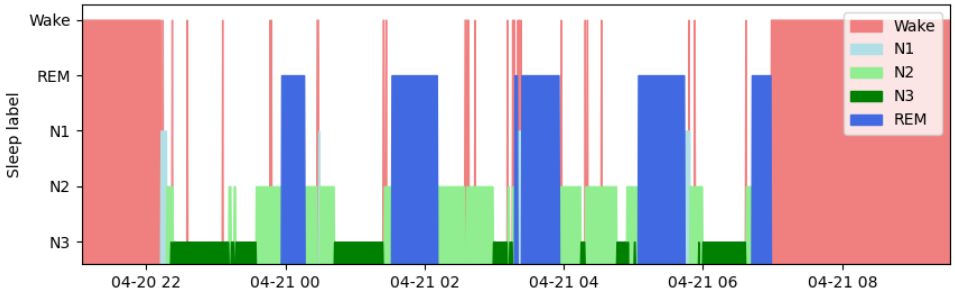



Figure 2.1. An example hypnogram showing sleep stages over a night's sleep.

**Physiological importance.** The different phases of sleep are associated with different restorative processes. For example, REM sleep is believed to play an important role in emotional processing [93]. Patients suffering from depression are known to have longer bouts of REM sleep that occur earlier during the night [27]. Meanwhile, deep (N3) sleep is important for memory consolidation and retention [48]. As we age, the amount of sleep we get (particularly of deep sleep) decreases, however, a greater decrease in deep sleep is associated with neurodegenerative diseases such as Alzheimer's disease [121].

In the clinic, quantitative sleep stage information, such as the total time spent in each stage, can be used to diagnose sleep disorders such as narcolepsy, parasomnias, and periodic limb movement disorders, and to aid in the diagnosis of broader conditions such as depression [115]. Quantitative analysis of sleep parameters including the sleep stages and their relative timings is carried out using a technique called polysomnography, which is described in the next section.

### 2.2.2 Polysomnography

Polysomnography (PSG) is a quantitative assessment of sleep quality. During a PSG recording, a wide variety of physiological signals are recorded over a night's sleep. In addition to the EEG and EOG (discussed in the previous section), these include: the electrocardiogram (ECG), which monitors the electrical activity of the heart; the chin electromyogram (EMG), which measures muscle activity of the chin; and respiratory inductance plethysmography (RIP) signals, from which

estimates of the breathing rate and respiratory effort can be derived. The signals recorded are logged centrally to a recording device which is worn on the chest of the subject, enabling synchronisation of the various time series signals. PSG studies can be performed in a sleep laboratory or at home, though in both cases this requires the careful attachment of the numerous electrodes required, by a trained sleep technician or physiologist.

The additional use of video data i.e. video polysomnography (vPSG) is deemed the “gold-standard" [207], and is recommended by the AASM [6], because it provides visual information to help a sleep physiologist to classify sleep stages[1] and is especially important for diagnosing sleep movement disorders.

An example PSG set-up is shown in Figure 2.2, and a typical equipment set-up is listed in Table 2.1. This particular equipment list was used as part of the collection of the Oxford Sleep Volunteers dataset [37], used later in Chapters 3 and 4.

Table 2.1. Polysomnography equipment used for studies in the Oxford Sleep Volunteers dataset [37].

| Monitoring technique | Description |
|---|---|
| Electroencephalography (EEG) | Electrodes, which measure brain activity at multiple sites, are attached to the scalp with dissolvable glue. |
| Chin electromyography (EMG) | Three electrodes are attached to the chin and measure chin muscle activity. |
| Electrooculography (EOG) | Two electrodes are attached near the eyes and measure eye movements. |
| Electrocardiography (ECG) | Two electrodes are attached to the chest and measure the electrical activity of the heart. |
| Inductance plethysmography | Bands are placed around the abdomen and thorax and measure movements of the chest wall. |
| Leg electromyography | Electrodes are attached to each leg and measure activity. |
| Pulse oximetry | A probe is attached to a finger to measure oxygen saturation levels. |
| Nasal airflow monitoring | A nasal cannula is placed inside the nostrils and measures airflow. |
| Actigraphy | An accelerometer within the recording device attached to the chest measures activity. |

[1] Particularly wakefulness.



Figure 2.2. **An example of a patient set-up for a polysomnography study**. Sensors visible include: electrooculogram (EOG) sensors, adjacent to the eyebrows; a nasal cannula, held in position by tape; electromyogram (EMG) sensors; two respiratory inductance plethysmography (RIP) bands around the abdomen and thorax; and a pulse oximeter, attached to the finger. Image by Kuyohong, licensed under CC BY-SA 4.0 from Wikimedia Commons.

**Sleep scoring.** After a night's sleep, a human expert (sleep physiologist) will manually review the recording. This process includes the annotation of sleep stages according to a set of scoring rules, usually those from the AASM. The scoring rules used to annotate the sleep stages require the physiologists to look for characteristic patterns in the recorded signals (primarily the EEG) such as sleep spindles [9] or slow delta waves [57]. Example signal characteristics during specific sleep stages are described in Table 2.2 and illustrated for the EEG in Figure 2.3.

Table 2.2. Polysomnogram characteristics of specific sleep stages.

| Sleep Stage | Sensor | Observed Behaviour |
|---|---|---|
| Wake | EOG | Blinking. |
| N1 | EEG | Disappearance of alpha rhythms [149].† |
| N1 | EOG | Slow, roving eye movements. |
| N2 | EEG | Sleep spindles and K-complexes [9].‡ |
| N3 | EEG | High-amplitude delta waves [57].† |
| REM | EMG | Diminished muscle tone. |
| REM | EOG | Rapid eye movements. |

†Alpha rhythms and delta waves are characteristic frequencies in the EEG signal.

‡Sleep spindles and K-complexes are characteristic EEG waveform shapes.

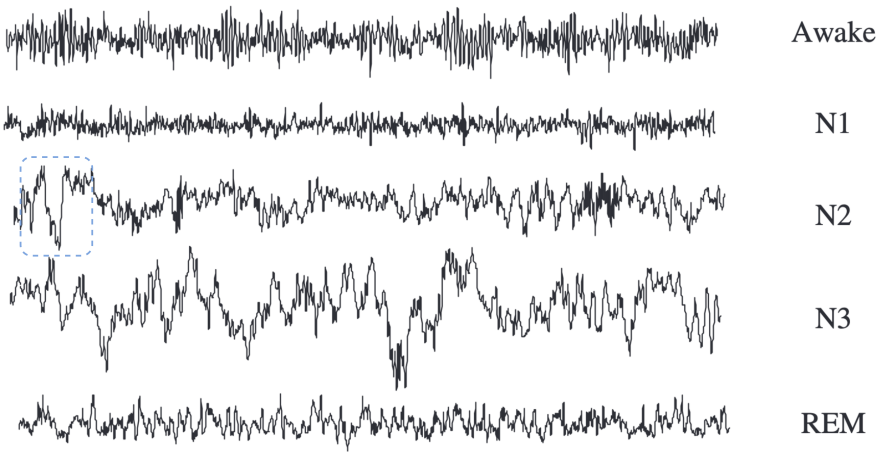



Figure 2.3. **Example presentation of different sleep stages in 30-second segments of an EEG recording**. Transitions from Wake to N1 result in the disappearance of 'alpha' rhythms (8-13 Hz activity) [149]. N2 is characterised by patterns in the EEG such as 'K-complexes' [9], an example of which is highlighted in the dashed blue box. N3 gives rise to high-amplitude, low-frequency (1-4 Hz) 'delta' waves. REM gives rise to similar EEG activity to wakefulness. Figure adapted from [104] with permission. Copyright 2013 Elsevier.

Sleep stage annotations can then be used to produce a sleep hypnogram (as in Figure 2.1), and to produce sleep metrics, which give a summary of a night's sleep. Common sleep metrics and their definitions are listed in Table 2.3.

Table 2.3. Definitions of common sleep metrics.

| Sleep Metric | Description |
|---|---|
| Time In Bed (TIB) | Time spent in bed during the night by the subject. |
| Total Sleep Time (TST) | Time spent asleep over the course of a night. |
| Sleep Efficiency (SE) | The ratio of TST to TIB. |
| Sleep Opportunity (SO) | Total duration of time available for sleep during a night. |
| Sleep Onset Latency (SOL) | The time taken from the start of the sleep opportunity to the onset of sleep. |
| Latency to Persistent Sleep (LPS) | The time taken from the start of the sleep opportunity to the first 10 minutes of persistent sleep. |
| Wake After Sleep Onset (WASO) | The total duration spent awake during the sleep opportunity after the initial onset of sleep. |
| REM Onset Latency (ROL) | The time taken from sleep onset to the onset of the first REM cycle. |

**Datasets.** The curation of large-scale PSG datasets has played a crucial role in furthering our understanding of sleep. The National Sleep Research Resource

[260] currently manages a number of large historical PSG datasets such as the Sleep Heart Health Study (5000+ participants, [170]) and the Multi-Ethnic Study of Atherosclerosis (6000+ participants, [44]). Collectively these datasets have enabled the discovery of sleep-related risk factors for health outcomes (e.g. [178]) and improved our understanding of how sleep varies with age [192]. These datasets have also supported the development of algorithms for the automated classification of sleep stages from the recorded signals [163], including those described later in this thesis. Additional information about these datasets can be found in Appendix B.

### 2.2.3 Limitations

**Subjectivity.** There is well-documented subjectivity between sleep physiologists in scoring a night's sleep from PSG recordings, with many studies having shown that a pair of scorers will, on average, agree on around 80% of sleep stage annotations for a single recording [25, 181, 255]. However, when multiple scorers are tasked with annotating the same night of data, Berthomier *et al.* [28] showed that the asymptotic agreement tends toward around 65% as the number of scorers increases.

**Cost.** Since sleep stages are typically annotated at 30-second intervals, an 8-hour recording requires around 1000 expert annotations. Important clinical events such as apnoeas, periodic limb movements are also usually annotated, to help build up a more comprehensive picture of sleep quality. As a result, the entire annotation process can take several hours to complete. PSG studies are a time-consuming and costly clinical procedure, which leads to longer waiting times for patients.

**Discomfort.** The full PSG equipment is unwieldy to set up and uncomfortable for patients, which often means that their recording is unrepresentative of their typical night's sleep. This is commonly referred to as the "first night effect" [3]. Because of this effect, the first night of data is often excluded from the analyses of PSG recordings [83].

These limitations of PSG additionally mean that longitudinal studies, i.e. regular monitoring over days or weeks, are rare. A non-invasive alternative method for monitoring (binary) sleep/wake cycles is actigraphy [26], which only requires a single

wrist-worn device. This makes it easier to perform than PSG, and enables recordings over greater durations e.g. several weeks. However, the number of sleep parameters that it is possible to measure using actigraphy is greatly reduced, and prior work has shown that the specificity of actigraphy in classifying sleep is low (0.329, [139]).

## 2.3 Deep learning

Across many problem domains such as language, vision and audio, deep learning methods have attained superior performance over alternatives such as classical machine learning and heuristic-based methods [67]. The increasing use of deep neural networks has also been accompanied by a general trend toward decreasing the number of pre- and post-processing steps and moving toward fully end-to-end learning systems. For example, in the visual domain, classical approaches to image classification included using Support Vector Machines (SVMs, [31]) applied to lower-dimensional, human-designed representations of image patches e.g. Histogram of Oriented Gradient features [51]. However, these methods are now significantly outperformed by modern deep neural networks that operate directly on pixels, surpassing human-level performance on benchmark datasets such as ImageNet [184]. This section describes important deep neural network architectures and the learning algorithms that can be used to train them, which are used extensively in this thesis.

### 2.3.1 Multi-layer perceptrons

The multi-layer perceptron (MLP, [182]) is one of the major original artificial neural network architectures. Since the discovery of the error back-propagation learning algorithm [183] in the mid-1980s, MLPs have been extensively used for supervised learning tasks such as regression and classification. An MLP takes an input vector $\boldsymbol{x} \in \mathbb{R}^d$ and applies a sequence of non-linear processing layers to produce an output

vector $\boldsymbol{y} \in \mathbb{R}^C$. The operations of an MLP can be defined as:

$$\boldsymbol{h}^{(0)} = \boldsymbol{x} \tag{2.1}$$

$$\boldsymbol{h}^{(l)} = \sigma_l(\boldsymbol{W}^{(l)}\boldsymbol{h}^{(l-1)} + \boldsymbol{b}^{(l)}) \quad \text{for} \quad l = 1, \ldots, L-1 \tag{2.2}$$

$$\boldsymbol{y} = \boldsymbol{W}^{(L)}\boldsymbol{h}^{(L-1)} + \boldsymbol{b}^{(L)} \tag{2.3}$$

where $\boldsymbol{W}^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$ and $\boldsymbol{b}^{(l)} \in \mathbb{R}^{d_l}$ are learnt, linear transformations[2] and $\sigma_l$ is a non-linear activation function, such as a sigmoid function (2.4) or a rectified linear unit (2.5) [145].

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{2.4}$$

$$\text{ReLU}(x) = \max(0, x) \tag{2.5}$$

For classification problems, the softmax function (2.6) is typically applied to the output of the final layer to convert the continuous, real-valued output vector $\boldsymbol{y}$ into a valid categorical probability distribution $\boldsymbol{p}$ where:

$$p_i = \text{Softmax}(y_i) = \frac{e^{y_i}}{\sum_{j=1}^{C} e^{y_j}} \tag{2.6}$$

### 2.3.2 Supervised learning

The Universal Approximation theorem states that a (sufficiently wide) MLP is capable of approximating *any* real-valued, continuous function over a compact set [89]. However, finding the optimal set of neural network parameters for a given task requires a learning algorithm. Supervised learning algorithms use labelled data to perform this task, i.e. the desired mapping is learnt from a dataset of input-output pairs:

$$\mathcal{D} = \{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \mid i = 1, 2, \ldots, N\} \tag{2.7}$$

Using such a dataset, deep neural networks are commonly trained by minimising an objective function $\mathcal{L}(\theta)$ using a variant of gradient descent:

$$\theta_{t+1} = \theta_t - \eta_t \nabla \mathcal{L}(\theta_t) \tag{2.8}$$

[2]Commonly referred to as the 'weights'.

where $\theta$ collectively denotes the set of parameters of the neural network, and $\eta_t$ (the learning rate) controls the step size for each parameter update.

For a fully differentiable neural network, the gradient with respect to any parameter can be calculated efficiently using the error back-propagation algorithm [183]. Mini-batch (stochastic) gradient descent i.e. using gradient estimates from a random subset of the dataset $\mathcal{D}$, has helped to scale these methods to very large datasets. Additionally, adaptive learning rate algorithms such as Adam [105] are commonly used because they often lead to faster convergence and improved training stability. Alternative optimisation methods such as evolutionary strategies have previously been used successfully [185], but generally lead to a combinatorial increase in computational requirements as the number of neural network parameters grows.

For classification problems, a popular choice for the objective function is the cross-entropy loss function (2.9), to minimise the difference between an output probability vector $\boldsymbol{p} = f_\theta(\boldsymbol{x})$ and a one-hot encoded vector[3] $\boldsymbol{y} \in \mathbb{R}^C$ containing the target label, averaged over the examples in the training dataset.

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(p_i) \tag{2.9}$$

### 2.3.3 Convolutional neural networks

Like an MLP, a convolutional neural network (CNN, [119]) typically contains many layers, with each layer alternating between a linear transformation of the input and element-wise non-linearity. However, whereas in an MLP each layer is fully connected i.e. each intermediate output is influenced by every input to that layer, in a CNN the linear transformation is a learnt convolution of the input.

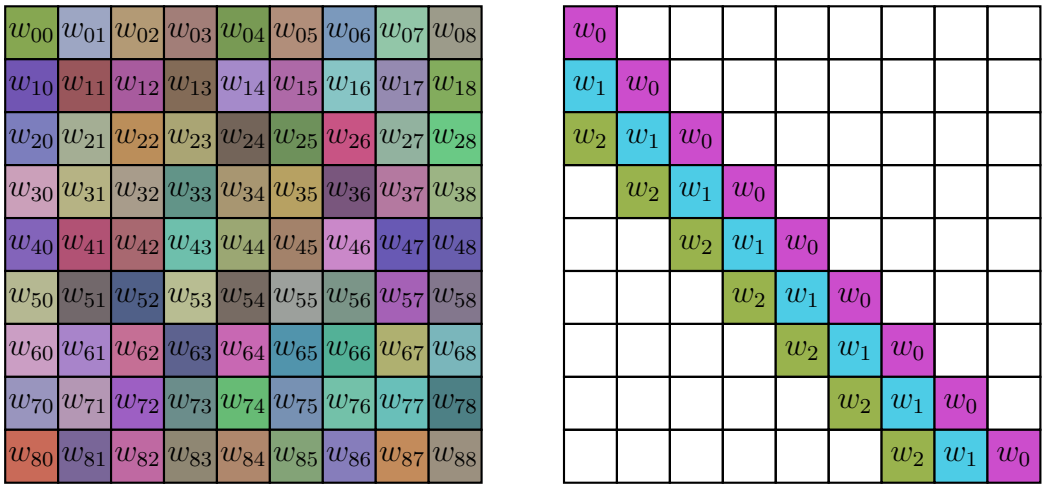
A 1D convolution with a single input and output channel can be expressed as:

$$y_n = \sum_{k=0}^{K-1} x_{n-k} \cdot w_k \tag{2.10}$$

Importantly, this means that only the inputs within a local receptive field determined by the kernel size $K$ influence each output $y_n$, and that the same transformation

[3]Where the element corresponding to the true class is equal to 1, and the rest are equal to 0.

is applied to every part of the input. This can alternatively be formulated as a sparse matrix-vector product with coupled weights, which is visualised in Figure 2.4. By extension of this formulation, for any given CNN, there exists an MLP that *could* represent the same function, by setting many of its weights to zero. However, for input domains where there is local structure, such as vision and time-series, the inductive biases[4] of CNNs mean that, in practice, they typically require fewer parameters and less training data to learn a desired mapping using techniques such as supervised learning.




(a) $\boldsymbol{W}_{\text{Linear}}$ (b) $\boldsymbol{W}_{\text{Conv}}$

Figure 2.4. Comparison of weight matrix structures for a linear transformation $\boldsymbol{W}_{\text{Linear}}$ and a 1D convolution $\boldsymbol{W}_{\text{Conv}}$ with kernel size K=3, expressed in the form $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$.

Ignoring bias terms, whilst a typical linear MLP layer $\boldsymbol{W}^{(l)}_{\text{Linear}}$ has $d_l \cdot d_{l-1}$ learnable parameters, a single-channel 1D convolutional operation only requires $K$ parameters. To improve model capacity, CNNs perform multiple convolutions within each layer. With $c_i$ input and $c_o$ output channels, this results in $c_i \cdot K \cdot c_o$ parameters for 1D input signals. This means that a CNN can use a much larger number of layers for the same number of total parameters and floating point operations per second (FLOPs) compared with an MLP. Architectural innovations such as depthwise-separable convolutions [45] can further help to reduce this parameter count.

Although each layer within a convolutional neural network only processes information within a local receptive field, stacking these layers progressively increases

[4]Locality and shift invariance.

the size of the effective receptive field. On tasks such as image classification, prior work has shown that CNNs learn progressively more abstract hidden representations with increasing network depth, with early layers learning operations that resemble low-level components of the visual cortex, such as Gabor filters [259].

Real-world image classification tasks require learning complex mappings from high-dimensional input images to output classes. Simonyan and Zisserman [198] showed that very deep convolutional neural networks could be used to learn such a mapping, achieving then state-of-the-art performance on various benchmarks with their 'VGG-Net' model. Residual neural network layers [78] have enabled further increases in network depth, by learning mappings of the form:

$$\boldsymbol{h}^{(l)} = \boldsymbol{W}_s^{(l-1)}\boldsymbol{h}^{(l-1)} + \mathcal{F}_\theta(\boldsymbol{h}^{(l-1)}) \tag{2.11}$$

within the sub-layers of the network. Here the linear projection $\boldsymbol{W}_s^{(l-1)}$ is used to account for any changes in the dimension between the input and output vectors induced by the arbitrary transformation $\mathcal{F}_\theta$. Empirically, He *et al.* [78] showed that this formulation enables stable training of convolutional neural networks with over 1000 layers, nearly two orders of magnitude greater than the 'VGG-Net' design, with no difficulty in optimisation [78]. These two design principles, very deep networks and residual layers, have been adopted for use in the vast majority of modern neural network architectures, including those used in the later chapters of this thesis.

### 2.3.4 Transformers

The Transformer [228] is a neural network design that was originally proposed for the sequence-sequence problem of language translation. However, in recent years it has been heavily used across many problem domains including vision [59], audio [15], and language [172]. This section describes the Transformer architecture and highlights the advantages which have led to its widespread adoption.

**Scaled Dot-Product Attention.** Transformer models are characterised by the use of an attention mechanism (2.12) to model relationships between sequences. Given sequences of input vectors $\boldsymbol{x}_{1:M} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\}$ and $\boldsymbol{z}_{1:N} =$

$\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N\}$, the scaled dot-product attention operation proposed by the original Transformer paper results in an output sequence of vectors $\boldsymbol{y}_{1:M} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M\}$, and can be formulated as:

$$\boldsymbol{y}_{1:M} = \text{Attention}(\boldsymbol{x}_{1:M}, \boldsymbol{z}_{1:N}) \tag{2.12}$$

$$\boldsymbol{y}_i = \sum_j \alpha_{ij} \boldsymbol{W}^V \boldsymbol{z}_j \tag{2.13}$$

$$\alpha_{ij} = \frac{e^{\tilde{\alpha}_{ij}/\sqrt{d_k}}}{\sum_j e^{\tilde{\alpha}_{ij}/\sqrt{d_k}}} \tag{2.14}$$

$$\tilde{\alpha}_{ij} = (\boldsymbol{W}^Q \boldsymbol{x}_i)^T (\boldsymbol{W}^K \boldsymbol{z}_j) \tag{2.15}$$

where $\boldsymbol{W}^Q \in \mathbb{R}^{d_\text{k} \times d_x}$, $\boldsymbol{W}^K \in \mathbb{R}^{d_\text{k} \times d_z}$ and $\boldsymbol{W}^V \in \mathbb{R}^{d_\text{y} \times d_z}$ are known as the query, key and value matrices. Each output vector $\boldsymbol{y}_i \in \mathbb{R}^{d_y}$ is a weighted sum of projections of $\boldsymbol{z}_j \in \mathbb{R}^{d_z}$ for all $j$, where the weighting is determined by the dot-product of projections of $\boldsymbol{x}_i$ and $\boldsymbol{z}_j$. The collection of weightings $\alpha \in \mathbb{R}^{i \times j}$, which quantify how much the projection of $\boldsymbol{z}_j$ contributes to output $\boldsymbol{y}_i$, is known as the attention matrix. The special case of attention where the two input sequences are the same i.e. $\boldsymbol{y}_{1:M} = \text{Attention}(\boldsymbol{x}_{1:M}, \boldsymbol{x}_{1:M})$ is known as 'self-attention', whilst attention between different sequences is commonly referred to as 'cross-attention'.

Notably, the attention operation (2.12) **does not depend on the position of vectors within the sequence**, and indeed Transformer models naturally operate on *sets* of vectors. For the Transformer to utilise positional information, an encoding is added to each vector in the sequence depending on its position, such as a learnt or sinusoidal bias [228].

**Multi-head Attention.** Rather than performing a single attention operation within each layer, a Transformer uses multiple 'heads' i.e. separate attention operations in parallel. By visualising the weights of the attention matrix, the original Transformer paper [228] showed that multi-head attention enables the model to perform different reasoning tasks simultaneously, which is necessary for language understanding. For example, one 'head' might specialise in identifying anaphora,[5] whilst another might focus on broader sentence structure. The outputs

[5]Reference to a prior contextual identity e.g. '*Ella* helped *Jonathan* tie *his* shoelaces'.

from $H$ heads are concatenated, meaning that each vector in the output sequence has a total dimension of $H \cdot d_y$. For simplicity, the number of heads and the output dimension from each head are usually chosen such that the resulting dimensionality of the output vectors is equal to that of the inputs i.e. $H \cdot d_y = d_x$.

**Transformer layers.** Building on the multi-head attention mechanism, there are three common Transformer layer designs, which are illustrated in Figure 2.5.

An encoder layer takes an input sequence of vectors $\boldsymbol{x}_{1:M} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\}$ and produces an output sequence of vectors $\boldsymbol{y}_{1:M} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M\}$ by alternating between multi-head self-attention over the set of vectors, and identical, non-linear transformations of each vector using an MLP.

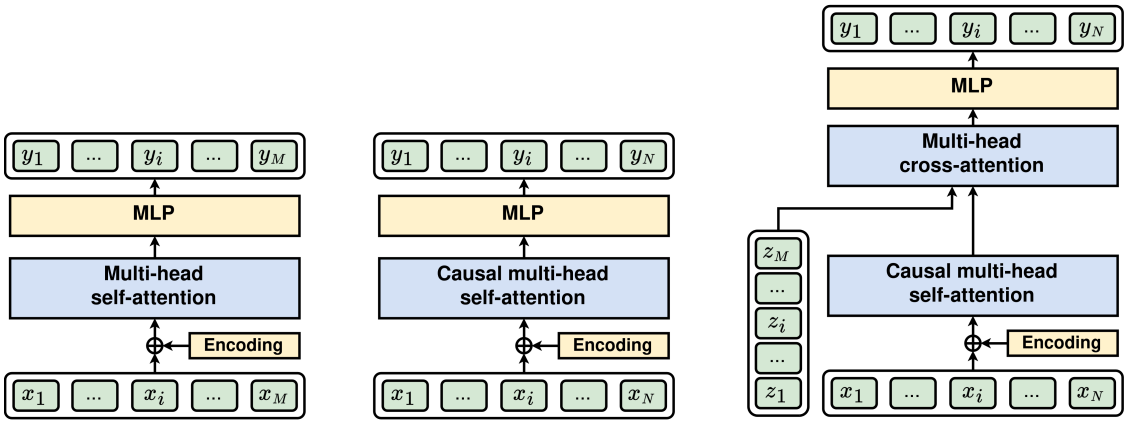
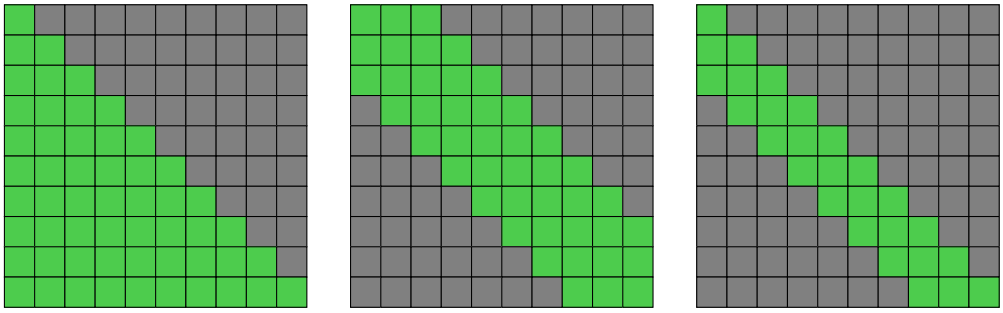



Figure 2.5. **Variants of the Transformer layer**. (a) Encoder-only layer. (b) Decoder-only layer. (c) Encoder-decoder layer. Element-wise operations are shown in yellow. Sequence-wise operations are shown in blue. Residual connections [78], dropout [204], and layer normalisation [14] are commonly used around both attention and MLP layers, but are omitted for brevity. Diagram inspired by the original Transformer design [228].

A decoder layer is a modification of an encoder layer such that the output vector $\boldsymbol{y}_j$ is only influenced by inputs $\boldsymbol{x}_{i \leq j}$. This is done by setting the un-normalized weights of the attention matrix $\tilde{\alpha}_{ij}$ (2.15) to $-\infty$ for $i < j$, so that the corresponding attention weights $\alpha_{ij}$ are equal to zero. This property is particularly valuable for causal tasks such as language generation, because each output element is produced using only the current and previous input elements for all positions in parallel.

Another common manipulation of the attention matrix is the use of local attention [24], which prevents each output sequence position from being influenced

by distant positions in the input sequence. Similarly to CNNs, stacking Transformer layers which use local attention progressively increases the effective receptive field. This can be a useful inductive bias for input sequences with hierarchical structure, allowing early layers to focus on short-range dependencies and later layers to combine global information. Additionally, local attention reduces the computational complexity of the attention mechanism from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$, the quadratic dependence being one of the major disadvantages of the original Transformer architecture. Common masks applied to the attention matrix are visualised in Figure 2.6.



(a) Causal attention (b) Local attention (c) Causal, local attention.

Figure 2.6. Visualisation of common masks applied to the Transformer attention matrix. Grey cells indicate $\alpha_{ij}$ is masked, i.e. element $i$ in the output sequence is not influenced by element $j$ in the input sequence (using row-major order indexing).

Finally, an encoder-decoder layer can be used to combine information from two input sequences to produce an output sequence, by using both self-attention and cross-attention. This design is particularly suited for tasks such as language translation where there is a source sequence (a sentence in one language) and a target sequence (a sentence in another).

**Composition.** Similarly to MLPs and CNNs, Transformer models use multiple layers to increase expressivity. Additionally, different Transformer layers can be composed together depending on the specific task. The original Transformer [228] first uses a stack of 6 encoder layers to turn the input sentence $\boldsymbol{x}_{1:M}$ into a source sequence $\boldsymbol{z}_{1:M}$. This is followed by a stack of 6 encoder-decoder layers which takes the source sequence and the output sentence $\boldsymbol{y}_{1:N}$ and predicts the next word in the output sentence for all output positions, i.e. the model is trained to predict $\boldsymbol{y}_{2:N+1}$ but, because of the causal attention mask, each output is only influenced

by previous words. After training, the model can be used to produce new output sequences auto-regressively, by repeatedly passing the input $\boldsymbol{x}_{1:M}$ and the existing output $\boldsymbol{y}_{1:k}$ to the model to predict $\boldsymbol{y}_{k+1}$ for all $k \leq N$.

**Advantages.** Because a Transformer decodes all output sequence positions in parallel during training, the operations can be significantly sped up using hardware accelerators such as Graphics Processing Units (GPUs). In contrast, recurrent architectures such as Long Short-Term Memory networks (LSTMs, [86]) traditionally required the entire input sequence to be processed *sequentially* before back-propagation can occur. The extensive use of parallelisation of a Transformer within the intermediate layers makes it well-suited for distributed training using multiple GPUs, which has led to the development of increasingly large models in recent years. For example, the GPT-3 Transformer model [34] has 175 billion trainable parameters. More recently, recurrent architectures such as Mamba [69] and xLSTM [23] have addressed the original limitations of architectures such as the LSTM, using efficient GPU implementations of parallel scan algorithms [30]. However, they are not yet as widely used as Transformers, which have extensive software support using libraries such as PyTorch [160], and have been successfully applied to a diverse range of applications.

As previously noted, positional information, such as the position of a word within a sentence, is typically injected into a Transformer using a manual or learnt bias vector [228]. This idea can be extended beyond 1D sequences. For example, in the vision domain, Transformer-based models typically split an image up into 2D patches, and use flattened (1D) image patches as the input vectors (tokens[6]) for the model, with 2D positional information added to each token [59]. This simple prerequisite, that the input needs to be 'tokenizable' has further improved the adoption of Transformers across different input modalities, and has simplified the development of models that can simultaneously operate on multiple input modalities, such as vision-language models e.g. [66, 213].

[6]Terminology adopted from the natural language processing (NLP) community.

In Chapters 4, 5 and 6 of this thesis, Transformers are used precisely for the reasons described in this section: they are effective for sequence-sequence modelling tasks; they can be easily scaled to multiple GPUs; and they can be used to combine data from multiple modalities.

### 2.3.5 Transfer learning

As discussed in Section 2.3.3, neural networks have been shown to learn progressively more abstract representations of the input with increasing network depth to solve a given task. For similar tasks e.g. object detection and image recognition, the necessary abstractions that a neural network must learn to solve each task are also similar. When training a neural network to solve a new task, the core idea of transfer learning is to reuse a model trained on a similar task (the source domain) and then fine-tune it on data for the new task (the target domain). Compared with training a new model from scratch i.e. with random initial parameters $\theta$, prior work has shown that this approach typically requires less data and less computation to achieve the same performance on the new task across numerous target domains [262].

**Frozen features.** Rather than fine-tuning a model on a new task, a popular variant of transfer learning is to instead freeze the model and use it as a fixed feature extractor [254]. In this set-up, the intermediate outputs (features) from one or more layers can then be used as inputs to downstream models. These intermediate features typically have lower dimensionality than the raw input and are chosen because they are known to encode useful information for the target task. By circumventing the need to learn a similar encoding using data from the target domain, this can help to mitigate overfitting and overcome data scarcity. In the image domain, this approach has been shown to significantly reduce the amount of labelled data required to reach a given level of accuracy [193]. An additional advantage of this approach is that the downstream model does not need to be a neural network, and can be trained via means other than error back-propagation. This means that alternative methods, which may be better suited to lower quantities

of data, can be used e.g. Random Forests [33]. In Chapters 3 and 4, a similar approach is used to address the problem of sleep staging from (scarce) video data.

## 2.4 Vision-based physiological measurement

A large body of prior work has shown that video cameras can be used to estimate measures of cardio-respiratory activity, including the heart rate and breathing rate [141]. Unlike traditional contact sensors such as the ECG, vision-based physiological measurement does not require skin preparation, does not cause skin irritation, and requires minimal equipment set-up [232]. This has motivated numerous studies which have demonstrated that vital signs such as the heart rate and breathing rate can be reliably measured in a clinical environment in a contact-free fashion, e.g. [217, 231, 233]. This section reviews the existing literature on the measurement of physiological parameters such as the vital signs (heart rate and breathing rate) and motion using video cameras. This underpins the work of Chapters 3 and 4, which builds on these advances to enable fully contact-free, vision-based sleep staging.

### 2.4.1 Cardiac monitoring

**Physiological basis.** Over the cardiac cycle, there are three physical phenomena detectable from a video camera from which the heart rate can be derived:

1. Blood volume changes, which are conventionally measured using a finger or wrist-worn optical sensor using a technique called photoplethysmography (PPG).

2. Colour changes, caused by the variation in blood oxygenation over the cardiac cycle and often referred to as "micro-blushes".

3. Head and body movements, caused by the ballistic forces generated as the heart pumps blood through the aorta. The measurement of these movements is known as ballistocardiography (BCG).

All three phenomena are typically imperceptible to the human eye, though techniques such as Eulerian Video Magnification [247] can be used to make the arising colour changes visible, as in Figure 2.7.



Figure 2.7. Eulerian Video Magnification of colour changes during the cardiac cycle. Reproduced from [247] with permission. Copyright 2012 ACM.

**Classical methods.** Camera-based remote photoplethysmography (rPPG) algorithms aim to estimate the pulse rate from a video camera by measuring the fluctuations in pixel values that arise from changes in blood volume. Poh *et al.* [168] showed that the cardiac signal can be extracted from the RGB channels of full-frontal facial video recorded using a webcam. They use Independent Component Analysis to find an optimal weighting of channels to separate the underlying signal from noise, which implicitly uses the fact that the relative strength of the rPPG signal varies with wavelength [242].

Balakrishnan *et al.* [19] measure the pulse from an RGB camera by measuring ballistocardiographic effects. They use the Viola-Jones detector [234] to identify the subject's face followed by the Lucas-Kanade tracker [135] to identify key points. Principal Component Analysis (PCA) is then applied to the set of time-series signals formed by measuring the movement of the feature points over time, to produce a single pulsatile signal which tracks the cardiac pulse. This enables the separation of motion due to the pulse from other sources present in the signals.

**Deep learning methods.** Rather than explicitly trying to measure specific physiological phenomena such as ballistocardiographic effects, improved accuracy has been achieved using deep learning-based approaches [141]. These use models that are typically trained using supervised learning to directly regress pulsatile signals

measured from contact sensors using video frames as inputs, e.g. DeepPhys [43], PhysFormer [257]. More recently, unsupervised deep learning methods have been proposed that can extract the cardiac pulse without supervision, by automatically discovering periodic signals within the video frames e.g. SimPer [253], SiNC [202].

### 2.4.2 Respiratory monitoring

Inhalation and exhalation result in quasi-periodic movements of the chest wall. These respiration-induced movements have traditionally been measured using contact sensors, with techniques such as inductance plethysmography (see Table 2.1), but they can also be measured using a video camera. Prior work has shown that this can be done through methods such as frame-differencing [16, 212], measuring dense optical flow magnitudes [146, 147], or by tracking the movement of sparse optical flow vectors in the video sequence [125].

### 2.4.3 Real-world use and validation

**Clinical studies.** Vision-based vital-sign monitoring has been validated in the clinic in a range of scenarios. Trumpp *et al.* [217] have monitored the heart rate of patients in an intra-operative setting, using a Bayesian classifier to identify the patient's skin from which to extract the rPPG signal. Meanwhile, Villaroel *et al.* [231], from the Oxford Biomedical Signal Processing & Machine Learning research group (BSP-ML), have measured the heart and breathing rate of patients undergoing haemodialysis on the ward, validating their algorithm on over 300 hours of data. More recent work from the BSP-ML group has proven the effectiveness of non-contact vital-sign monitoring in the neonatal intensive care unit [233] and in the adult intensive care unit for post-operative monitoring [98].

**Overnight monitoring.** Vision-based physiological measurement has also been validated during overnight vPSG recordings, both in healthy young adults [238] and in subjects with suspect obstructive sleep apnoea [223]. This has been done using cameras that operate outside the visible spectrum, e.g. using near-infrared (NIR) cameras [244]. The relative signal strength of the cardiac pulse is typically

much weaker using these wavelengths [243]. However, multi-channel NIR set-ups [222] and custom-built light sources [244] have both been used to overcome this.

**Commercial applications.** Oxehealth's Oxevision$^{TM}$ software for non-contact vital-sign monitoring has received Class IIa medical device certification from the Medicines and Healthcare products Regulatory Agency (MHRA) in the UK and De Novo clearance from the US Food and Drug Administration (FDA) [219]. The system is deployed in over three thousand rooms across clinical settings including elderly care and mental health hospitals, where it has been shown to lead to a better night's sleep for patients [130]. In Chapters 3 and 4, this software is used to extract cardiac and respiratory signals from video data.

### 2.4.4 Video actigraphy

The aim of video-based actigraphy methods is to derive a continuous measure of activity directly from a video camera, and these methods have been used for tasks such as binary sleep-wake classification [132]. This is particularly useful in settings such as neonatal care, where wearable devices are unsuitable [131].

Video actigraphy signals have previously been extracted using methods such as counting pixel differences above a threshold [191] and spatial averaging of optical flow vectors [49, 147]. Whilst conventional wrist-based accelerometry provides a measure of activity from a single point on the body, an additional advantage of video-based actigraphy is that it can be used to calculate multiple activity signals from specific body regions simultaneously, e.g. to detect leg movements [238].

### 2.4.5 Future additions

Though not explored in this thesis, there are other physiological parameters that could be extracted from a near-infrared camera, and which could in future be used to aid the classification of sleep stages.

**Vital signs.** Skin temperature can be continuously measured using thermal cameras [128]. The estimation of blood oxygen saturation [70] and blood pressure [90] from video cameras has also been explored, though the accuracy of these

techniques lags behind that of contact sensors [76]. Continuous measurement of blood oxygen saturation could additionally be used to detect sleep apnoeas [235].

**Sleep pose.** Mohammadi *et al.* [143] have explored classifying sleep posture from video. Whilst there is no direct link between specific sleep posture and sleep stages [133], the variation (or a lack thereof) in sleep pose over time may be predictive of the sleep stage. For example, given that normal REM sleep induces temporary paralysis, we would not typically expect sleep pose to change during this period.

**Visual semantics.** Finally, as discussed in Section 2.2.2, human experts use the video data from a vPSG recording to aid the classification of sleep stages during obvious wakeful periods and to identify abnormal movements during sleep. Visual semantics, such as whether the subject is reading a book or using a phone, can be particularly helpful for human experts to identify periods of 'quiet wakefulness', for which techniques such as actigraphy are known to struggle [26]. Visual question-answering models [10] could enable this semantic information to be automatically extracted and used by an automated sleep staging algorithm.

## 2.5 Automated sleep staging

The goal of automated sleep stage classification (sleep staging) algorithms is to produce sleep stage time-series outputs $y_{1:T}$ from one or more input modalities, such as the signals recorded during a PSG study. Sleep staging algorithms are typically trained and evaluated against AASM (5-class) annotations of PSG recordings by human experts. However, given the low inter-expert agreement in identifying N1 [52], many works (e.g. [110, 174]) have converted this to a 4-class problem, where N1 and N2 are combined into a single 'Light' sleep class, or into a 3-class problem (e.g. [209]), where N1, N2 and N3 are combined into a single 'NREM' class.

**Evaluation metrics.** In addition to multi-class classification accuracy, sleep staging algorithms are commonly evaluated using Cohen's $\kappa$ statistic [46]. This is a measure of inter-rater agreement, with a value of 0 or less indicating no agreement and a value of 1 indicating perfect agreement between sources. Cohen's $\kappa$ statistic is more robust to class imbalance, which is common in the domain of sleep staging,

Table 2.4. A rule of thumb for assessing Cohen's $\kappa$ [118].

| Value | Strength of Agreement |
|---|---|
| $\leq$0.00 | Poor |
| 0.00-0.20 | Slight |
| 0.21-0.40 | Fair |
| 0.41-0.60 | Moderate |
| 0.61-0.80 | Substantial |
| 0.81-1.00 | Almost perfect |

and accounts for the probability of two sources agreeing by chance. Additionally, its usage explicitly acknowledges the fact that, as discussed in Section 2.2.3, expert-annotated sleep stages are subjective and *not* a ground truth. A commonly used rule of thumb[7] for assessing Cohen's $\kappa$ is given in Table 2.4.

**Modalities.** Many prior works have addressed the problem of automated sleep staging, and these have used a wide range of combinations of input modalities. The remainder of this section divides prior work on automated sleep staging into three categories:

- PSG-based sleep staging. Works that rely on signals typically recorded as part of a polysomnograhy study, including EEG, EMG, and/or EOG signals.
- Wearable sleep staging. Works that have explicitly used modalities that are obtainable from a wearable device or other contact sensors e.g. PPG, ECG and/or accelerometry signals.
- Non-contact sleep staging. Works that have used a contact-free sensor, such as a video camera, as their only input source.

Given the drawbacks of polysomnography discussed in Section 2.2.2, the last two categories of methods hold significant advantages. They are simpler and cheaper to administer, and provide a more comfortable experience for patients. In turn, this can reduce the cost of validating sleep disorder treatments, through cheaper and easier clinical trials, and enable longitudinal sleep monitoring over multiple

[7] Often incorrectly attributed to Cohen himself.

days or even weeks. Improving the utility of sleep staging from these last two categories of modalities is the primary focus of this thesis.

### 2.5.1 PSG-based sleep staging

The EEG is one of the primary modalities used by sleep physiologists when classifying sleep, and a recent literature review has claimed that automated sleep staging using EEG data has effectively been solved [163]. In other words, with access to the same modalities as human scorers from a PSG recording, state-of-the-art machine learning models can classify sleep stages with model-expert agreement comparable to that between two human experts. For example, the SleepTransformer model [166] attains an accuracy of 87.7% and a $\kappa$ of 0.83 in full 5-class (AASM standard) sleep staging on the Sleep Heart Health Study (SHHS) dataset. The model uses EEG time-frequency maps for each sleep epoch as inputs and a transformer encoder [228] to combine information temporally, jointly classifying adjacent sleep stages.

### 2.5.2 Wearable sleep staging

Although human experts primarily rely on EEG characteristics to classify sleep stages, correlates of the sleep stages can also be observed in cardiac, respiratory and activity signals. For example, respiration tends to be more irregular during REM sleep [91], whilst low-frequency power (0.04-0.15Hz) in the heart rate time series is typically lowest during slow-wave (N3) sleep [195]. This has led several works to investigate classifying the sleep stages using only cardio-respiratory signals, e.g. ECG and respiratory signals [179], which can be less obtrusive to acquire than EEG signals.

Prior work has also investigated the potential of classifying sleep stages using wrist-worn wearable devices, such as smartwatches. These devices are typically able to measure the heart rate, from PPG or ECG signals. Respiratory rate measurement is also possible from these sensors since respiratory activity causes amplitude modulation of both the ECG and PPG signals [41], although this phenomenon decreases with the age of the subject, see for example [152].

Accelerometry can further aid the classification of sleep stages, as there is a strong relationship between the rate of body movements and the sleep stage [246]. Alternatively, it can be used as a stand-alone modality from which to perform sleep staging [186]. In a 2023 white paper [11], Apple reported that their proprietary sleep staging algorithm for the Apple watch obtains a Cohen's $\kappa$ of 0.63 in four-class sleep staging, grouping N1 and N2 into a single class. This uses 3-axis accelerometry signals recorded by the watch which, in addition to gross body movements, can track subtle movements caused by breathing.

As seen with PSG-based sleep staging, deep learning-based methods have become the dominant approach to sleep staging from non-EEG sensor combinations such as the ECG [173, 203], ECG and respiratory signals [17, 209], and PPG signals [110, 174]. These have achieved significantly higher accuracy than classical machine learning methods, e.g. [64, 179]. Radha *et al.* [174] used a transfer learning approach to classify sleep stages from a wrist-worn PPG sensor. The authors first train an LSTM [86] on the SIESTA dataset [106], using time-series features derived from the ECG signal as inputs. The model is then fine-tuned using identical features derived from the PPG signal on a smaller dataset, achieving an accuracy of 76% ($\kappa = 0.65$). Their approach highlights how large PSG datasets can be utilised to improve the classification performance of sleep staging models which use data-scarce modalities such as the PPG. More recently, Kotzen *et al.* [110] showed that using minimally-processed PPG time series as the sole input to a CNN, rather than time-series features, can lead to improved performance, mirroring the wider trend toward end-to-end deep learning methods discussed in Section 2.3.

### 2.5.3 Non-contact sleep staging

**Video-based sleep staging.** As discussed in Section 2.4, vision-based physiological measurement of parameters such as the heart rate and breathing rate has been extensively explored in the literature [141], validated in clinical settings [98, 233], and led to FDA-cleared commercial software for physiological measurement [219]. This raises the possibility that sleep staging could alternatively performed in a

completely contact-free manner using physiological signals measured entirely from a video camera, further simplifying the process of overnight sleep monitoring. Within the field of sleep medicine, the collection of video data is common practice as it enables the accurate diagnosis of specific conditions such as parasomnias and REM behaviour disorder [4, 206]. However, given its sensitive nature, collection and use of video data can only take place with patient consent and clinical assessment. Beyond traditional sleep medicine, video-based monitoring may also have advantages over wearable-based methods in certain healthcare settings. For example, in elderly care, e.g. dementia wards, video data can simultaneously be used for detecting or predicting falls, one of the leading causes of death amongst older adults.

Only a few prior works have explicitly attempted to classify sleep into Wake, NREM and REM phases entirely from video data i.e. beyond binary sleep–wake classification. Nochino *et al.* [153] used a video actigraphy signal to derive a number of motion features for each sleep epoch and trained a model to classify sleep stages using these features. However, their model achieves a Cohen's $\kappa$ of only 0.19 using 24 nights of data from 6 volunteers, highlighting that activity measurements alone are likely insufficient to classify sleep stages. Using similar methods, Kamon *et al.* [100] attained similar classification performance on a dataset of vPSG recordings from 8 infants.

van Meulen *et al.* [226] use a camera-derived pulse signal to perform video-based sleep staging. From the pulse signal, they derive several heart rate variability (HRV) features which are passed to a neural network to classify sleep stages, obtaining a $\kappa$ of 0.49. The neural network is trained using HRV features derived from a separate contact-sensor dataset. The model is not fine-tuned or otherwise adapted using the video data, which could further improve performance. Additionally, their method does not make use of motion or respiratory signals, which (as detailed in Section 2.4) can be robustly measured from video data during sleep, and which are also predictive of the sleep stage.

**Alternatives.** Non-contact sleep monitoring has also been investigated using devices such as radar [87, 116, 216], microphones [50, 252], and mattress-based

sensors [58, 200], which are similarly capable of extracting cardiac, respiratory and/or motion signals. Given the extensive clinical evaluation and the real-world implementation of vision-based physiological measurement discussed in Section 2.4.3, it is arguably a more mature technology. However, the algorithms introduced later in this thesis would be applicable to alternative non-contact modalities which may be better suited to some clinical settings.

### 2.5.4 Foundational assumptions

**Discretisation.** It is important to remember that sleep stages are a (discrete) *human-defined* clustering of the (continuous) state of the body during sleep, which are a simplification of the behaviour we observe. For example, REM sleep is treated as single homogeneous physical state by the AASM scoring rules. However, REM can be further divided into phasic and tonic microstates, which have very different neural characteristics [199]. Additionally, research has shown that, rather than being in a uniform state, different parts of the brain can be in different stages of sleep simultaneously [112, 196]. Despite their coarse definitions, discretised sleep stages have proven their utility for diagnosing sleep disorders, identifying biomarkers, and validating sleep disorder treatments.

**Label shift.** Sleep scoring guidelines such as those of the AASM continue to slowly evolve alongside our understanding of sleep physiology, with occasional amendments and updates. These are typically minor, such as additional guidance on identifying (N3) slow-waves [7]. However, the AASM scoring rules have been critiqued [188, 205] and it is possible that, at some point in the future, there may be a further change in the stratification of sleep stages, similar to the original change from the R&K to the AASM standard discussed in Section 2.2.1. Under the supervised deep learning paradigm, which has to-date achieved state-of-the-art results in automated sleep staging [163], this would likely require re-training of models with sleep stage annotations against any new standard. If this were to occur, a transfer learning approach using existing datasets and annotations could reduce the necessary

amount of data annotated according to any new standard, as has previously been done in the transition from R&K to AASM-annotated sleep stages [174].

## 2.6 Conclusions

The conventional “gold-standard" method for sleep monitoring is uncomfortable for patients, unwieldy for technicians to administer, and unsuitable for long-term monitoring over many days or weeks. Additionally, there is a high degree of subjectivity in the classification of sleep stages by human experts. Deep learning methods have enabled automated sleep stage classification on a par with human scorers using EEG data, but this uses electrodes attached to the scalp, maintaining many of the drawbacks of traditional PSG. Using wearable devices, which measure surrogates such as cardio-respiratory signals, other prior work has investigated sleep stage classification with a much less obtrusive set of input modalities, albeit obtaining lower accuracy. Improving the accuracy of wearable-based methods could make them viable for applications such as longitudinal, at-home sleep monitoring. Finally, from a video camera it is possible to measure cardio-respiratory signals and activity levels during sleep without the need for any sensors to be attached to the individual being monitored. This opens up the possibility of accurate sleep stage classification in an entirely contact-free fashion, which is explored in the next two chapters.

# Part I

# Sleep Staging from Vision

# 3

# Video-based Sleep Staging using Transfer Learning

## Preface

This chapter contains the paper “Deep Learning-Enabled Sleep Staging From Vital Signs and Activity Measured Using a Near-Infrared Video Camera". This was presented at the 6th Workshop on Computer Vision for Physiological Measurement at CVPR 2023, where it received an Honourable Mention for Best Paper award. I was the lead author of the paper, devised the proposed transfer learning method, conducted all of the experiments, and wrote the initial manuscript. A full Statement of Authorship can be found in Appendix A.

The enclosed paper was the first to show that multiple modalities from video (heart rate, breathing rate[1] and motion) could be combined to achieve video-based sleep staging with high agreement against expert-annotated sleep stages, outperforming the few prior works that had addressed this problem using a single video-derived modality.

[1]The terms ‘breathing rate’ and ‘respiratory rate’ are used interchangeably in this thesis.

## Contents

**3.1 Introduction . . . . . 37**
**3.2 Related work . . . . . 40**
**3.3 Methods . . . . . 41**
3.3.1 Data acquisition . . . . . 41
3.3.2 Optical flow-based activity measurement . . . . . 42
3.3.3 Motion features . . . . . 43
3.3.4 Video-based vital-sign monitoring . . . . . 44
3.3.5 Vital-sign representation learning . . . . . 45
3.3.6 Transfer learning . . . . . 48
3.3.7 Evaluation . . . . . 49
**3.4 Results and Discussion . . . . . 49**
3.4.1 Ablation studies . . . . . 52
**3.5 Conclusions . . . . . 52**
**3.A Chapter Appendix . . . . . 53**
3.A.1 Vital-sign estimation from contact sensors . . . . . 53
3.A.2 Further model implementation details . . . . . 54
3.A.3 Feature engineering baseline . . . . . 54
3.A.4 Additional model hypnograms . . . . . 54

# 3.1 Introduction

The 'gold-standard' method for sleep monitoring is video polysomnography (vPSG, [207]), a test in which a number of physiological signals are recorded over a night's sleep. This typically requires an extensive set of contact sensors to be attached to the patient, which often leads to patient discomfort and an unrepresentative night's sleep [3]. After the recording, a human expert (sleep physiologist) must review the outputs from these sensors, manually annotating the record with sleep stages, leg movements and other events of interest from throughout the night. This process can often take multiple hours to complete. The cost of the study and the discomfort to the patient mean that longer-term studies are rare. Easier longitudinal monitoring would also benefit clinical research, from making trials for sleep-disorder treatments easier, to detecting the onset of dementia, by monitoring variations in slow-wave sleep [195].

Sleep stages are commonly annotated at 30-second intervals using the American Academy of Sleep Medicine (AASM) guidelines. These divide sleep into five discrete

states: Wake, N1, N2, N3 and REM sleep. N1 to N3 account for varying degrees of non-rapid eye movement (NREM) sleep from near-wakefulness (N1) to deep slow-wave sleep (N3).

There is well-studied subjectivity between sleep physiologists in scoring sleep stages, which is often reported using the Cohen's kappa statistic ($\kappa$, [46]). The work of Danker-Hopfe *et al.* [52] showed that pairs of scorers agreed on 82% of epoch-by-epoch sleep stages ($\kappa = 0.76$) across a dataset of 72 sleep recordings.

A recent review has claimed that deep learning methods can score PSG records as accurately as a human scorer [163]. The SleepTransformer model [166] achieves a Cohen's $\kappa$ of 0.83 in AASM sleep stage classification on the Sleep Heart Health Study (SHHS, [170]) dataset from an electroencephalogram (EEG) signal. However, the EEG is typically measured using electrodes which must be glued to the subject's scalp by a sleep physiologist.

To overcome this need, other work has investigated using cardio-respiratory and activity information obtained from wearable devices [167, 174, 236, 248]. Wearable sleep staging is possible from these inputs since they encode information about the underlying sleep state. For example, respiration tends to be more irregular during REM sleep [91], whilst low frequency power in the heart rate time series is typically lowest during N3 sleep [195]. Wearable-based methods have commonly simplified sleep staging into a four-class classification problem, by merging N1 and N2 into a 'light sleep' class. Doing so, Radha *et al.* [174] achieve a Cohen's $\kappa$ of 0.65 in four-class sleep staging from wrist-worn photoplethysmography.

Prior work has shown that both cardio-respiratory [223] and activity information [131] can be derived from video data during sleep, indicating that video could be used as a single source from which to classify sleep stages. As an entirely non-contact solution, this would further simplify and improve the sleep monitoring process for both patients and clinicians. Additionally, video remains an important modality for the diagnosis of conditions including periodic limb movement disorder, parasomnias and REM behaviour disorder [55, 206]. It may not be feasible to identify these conditions using alternative solutions such as wrist-worn wearables.

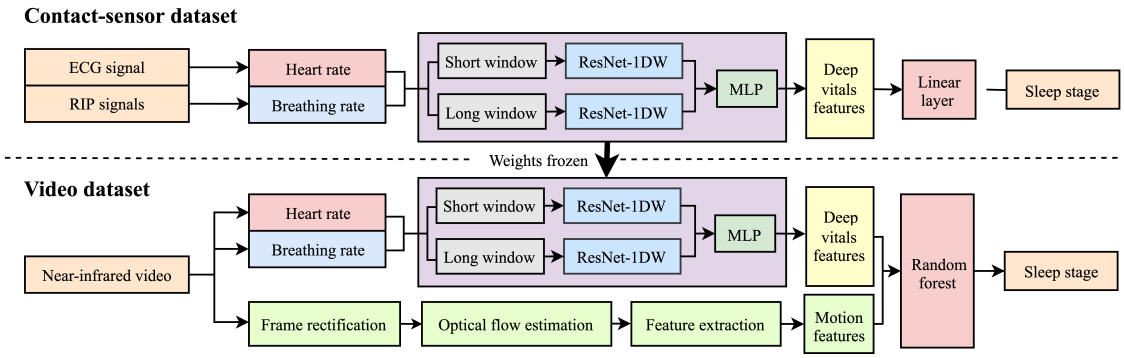



Figure 3.1. Overview of our approach. We first train a deep learning model on a contact sensor dataset to classify sleep stages from heart rate and breathing rate time series. We then use this trained model as a feature extractor to produce 'deep' features from the video-derived vital signs of our target dataset. To classify sleep stages from video data, we train a random forest classifier which uses these deep features in conjunction with motion features derived from optical flow.

In this work, we use heart rate, breathing rate and activity signals, all derived from near-infrared video, to perform sleep stage classification. We use optical flow [88] to derive continuous measures of activity during sleep, which are made robust to camera position and bed orientation using image homography transformations [77].

To overcome data scarcity, we adopt a transfer learning approach (Fig. 3.1). We first train a deep neural network to classify sleep stages from heart rate and breathing rate time series derived from a substantially larger contact sensor dataset. The trained model is then applied to heart rate and breathing rate time series derived from near-infrared video to obtain 'deep' features. These deep features are used in combination with video-derived activity features as inputs to a classifier which is trained and evaluated on a video dataset of single-night data from 50 healthy volunteers using 10-fold cross validation.

Using our approach, we present improved results over the existing state-of-the-art [226] for video-based four-class sleep staging in terms of both accuracy (73.4% vs. 67.9%) and Cohen's kappa (0.61 vs. 0.49).

## 3.2 Related work

**Video actigraphy:** Counting pixel differences above a threshold has proven to be a simple but effective activity signal for binary sleep-wake classification from both near-infrared (NIR, [191]) and depth images [113, 229].

Optical flow [49, 147] and spatio-temporal recursive search algorithms [80] have also proved effective at quantifying activity. These methods both compute motion vectors from the video feed and use average magnitudes as activity signals. The latter method has been used to detect periodic limb movements during sleep [81].

Using activity features derived from video, Long *et al.* [131] achieve an accuracy of 92.0% in sleep-wake detection from video using a dataset of 10 healthy infants.

**Video-based sleep staging:** To the best of our knowledge, only two prior works have attempted to classify sleep beyond a binary sleep-wake classification using a video camera. These works differ greatly in the information which they extract from the video to classify sleep stages.

Nochino *et al.* [153] use frame-differencing to derive an activity signal from the camera. From this signal, the authors derive a number of motion features for each sleep epoch and train a Support Vector Machine (SVM) to perform four-class sleep staging, achieving an average accuracy of 40.5% and a Cohen's $\kappa$ of 0.19 on a dataset of 6 participants.

van Meulen *et al.* [226] also classify sleep stages into four classes, but do so using camera-derived inter-pulse intervals (IPIs). The authors directly apply a deep learning model, previously trained on heart-rate variability (HRV) features derived from the electrocardiogram (ECG), to HRV features computed from a camera-derived IPI time series. This achieves an accuracy of 68% and a Cohen's $\kappa$ of 0.49 across 46 healthy participants.

## 3.3 Methods

### 3.3.1 Data acquisition

In this work, we use video as the input data and polysomnogram recordings as the source of sleep stage labels collected from 50 healthy adult volunteers (18+) with no previous diagnoses of sleep disorders. We report age, weight, sex and Fitzpatrick skin type for the study population in Table 3.1. Fitzpatrick skin type [63] is a measure of skin pigmentation which classifies skin phototypes into one of six categories ranging from I (very pale) to VI (very dark).

Table 3.1. Sleep study population demographics.

| Variable | Value |
|---|---|
| Age[1] | 37±15 years |
| Weight[1] | 70.4±12.9 kg |
| Sex | |
| Male | 22 volunteers |
| Female | 28 volunteers |
| Fitzpatrick skin type[2] | I:13, II:21, III:12, IV:3, VI:1 |

[1]Mean (± std). [2]Skin type: No. volunteers.

Each recording took place overnight in one of two experiment rooms. Video data for automated sleep staging was captured in each room at 20 FPS using a single-channel 850 nm NIR video camera. Both rooms also contained LED illuminators centred at the same wavelength and placed adjacent to the cameras. Figure 3.2 shows the most common bed orientations and the fixed camera positions used in each room.

Video polysomnography data was acquired using a SOMNOscreen$^{TM}$ Plus device, which recorded a typical vPSG montage which included respiratory inductance plethysmography (RIP), electrooculogram (EOG) and EEG signals and video data, using a separate camera set-up.[2] Each recording was manually scored according to AASM guidelines in 30-second windows (sleep epochs) by an experienced sleep physiologist.

[2]The camera included with the vPSG kit, used by the sleep physiologist for their analysis.

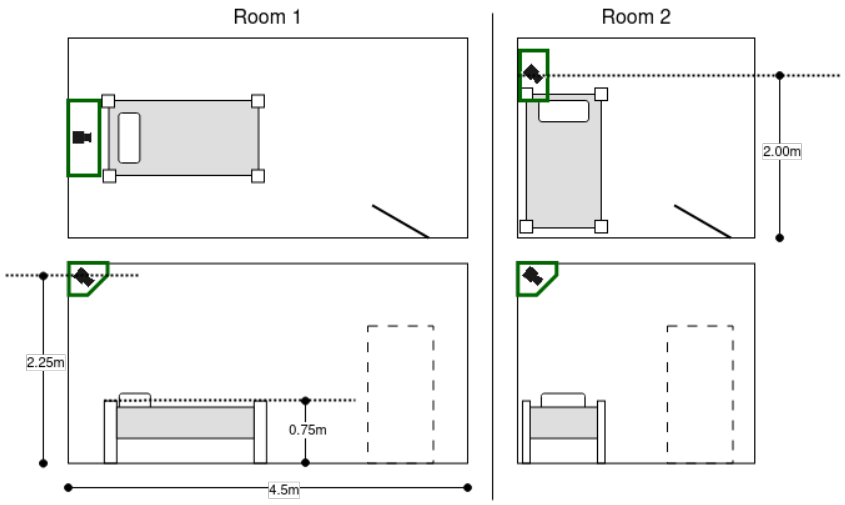



Figure 3.2. Room and camera layouts used for data acquisition.

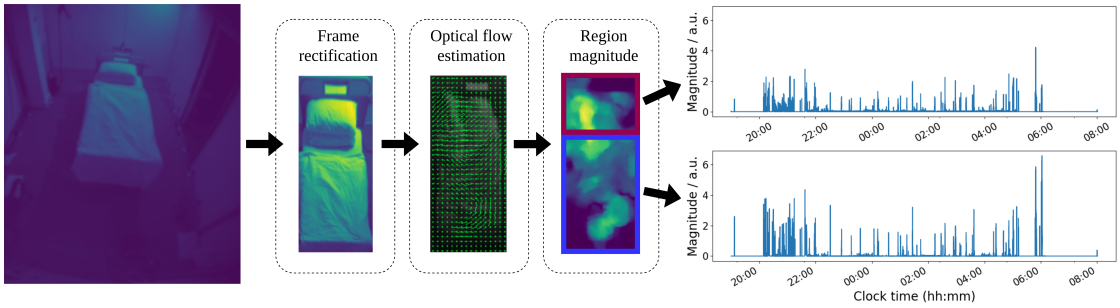



Figure 3.3. Extracting 1D activity signals from video using optical flow. A homography transformation is applied to each frame followed by cropping to the bed region, to obtain a consistent, virtual viewpoint. The magnitudes of optical flow vectors calculated between frames are averaged over the upper and lower bed regions to produce continuous activity signals for each region.

### 3.3.2 Optical flow-based activity measurement

From the video camera, we derive continuous measures of subject activity using optical flow [88]. This process is illustrated in Figure 3.3. First, we apply an image homography transformation [77] which provides a consistent, virtual viewpoint of the bed region for different camera positions. This technique has previously been used to improve the robustness of sleep pose detection [144]. We then compute optical flow between transformed frames using the Dense Inverse Search algorithm [111]. From the 20 FPS video, we estimate the optical flow field between every 5th frame, then average over the previous four estimated fields, giving a field estimation frequency of 4 Hz.

From the optical flow field time series, we derive two simple activity signals by averaging the flow field magnitude in the upper third and lower two thirds of the rectified view, which correspond to approximate head and lower body regions. This gives a simple measure of the spatial scale of movements, since whole-body movements such as sleep posture changes will result in high activity for both signals.

### 3.3.3 Motion features

From our two video-derived activity signals, we compute a total of 20 time-series features for each sleep epoch, listed in Table 3.2. These features broadly fall into two categories: integral features, which quantify the amount of motion that occurred within a time window centred on the epoch, and counter-based features, which quantify the elapsed time since a movement occurred. Our counter-based features are inspired by those of Nochino *et al.* [153].

Table 3.2. Motion features derived from the 1D optical flow activity signals for each sleep epoch. Features are computed from the upper and lower bed region signals separately.

| Count | Feature description |
|---|---|
| **14** | **Integral features** |
| 6 | 30-second Gaussian-weighted sums, time-shifted by (-30, 0.0, 30.0) seconds. |
| 6 | 180-second Gaussian-weighted sums, time-shifted by (-180.0, 0.0, 180.0) seconds. |
| 2 | 20-minute Gaussian weighted sums. |
| **6** | **Counter-based features** |
| 4 | Time elapsed since activity signal was above a threshold $\in \{0.1, 1.0\}$. |
| 2 | Time elapsed since activity signal was above a threshold (1.0) for 3 seconds. |

### 3.3.4 Video-based vital-sign monitoring

Many prior works have shown that both the heart rate [124, 168, 230, 241] and breathing rate [125, 147, 240] can be measured using video cameras. These methods have been validated in a number of clinical settings e.g. [98, 217, 233]. The recent study of van Gastel *et al.* [223] has also demonstrated the feasibility of continuous vital-sign monitoring during sleep.

In this work, we use components of Oxevision Vital Signs medical device software [92] to measure the heart rate and breathing rate from video data. For both vital signs, this produces 1 Hz estimates and binary Signal Quality Indices (SQIs, [126]), which indicate the software's assessment of the estimate's quality.

The software estimates both vital signs through a process of region of interest (ROI) selection, spatio-temporal filtering, cardio-respiratory signal extraction and data fusion. First, ROIs are identified by selecting areas in the image containing high variance that is periodic in nature [92]. These ROIs are then spatially smoothed with a median filter and band-pass filtered using pass-bands of 40–130 beats per minute (bpm) and 8–39 breaths per minute (brpm) for the heart and breathing rates respectively. The cardio-respiratory signals are extracted from the ROIs using principal component analysis as done in Villaroel *et al.* [233], with the heart and breathing rates identified from the dominant frequency components. SQIs are derived for each vital sign from heuristics including inter-beat interval consistency. These SQIs are typically low when the volunteer is active, e.g. out of bed, or absent from the room.

We define the coverage of each vital sign as the percentage of time for which the SQI is high when the volunteer is in bed. Using this definition, Figure 3.4 shows the coverage for each vital sign across recordings in the dataset. Coverage statistics computed across participants are given in Table 3.3.

Figure 3.5 shows an example of heart rate and breathing rate time-series estimated from near-infrared video over a night's recording along with AASM-annotated sleep stages. In this example, we can observe increased variance in the

breathing rate around REM cycles, indicating how sleep state information can be encoded in the cardio-respiratory time series.

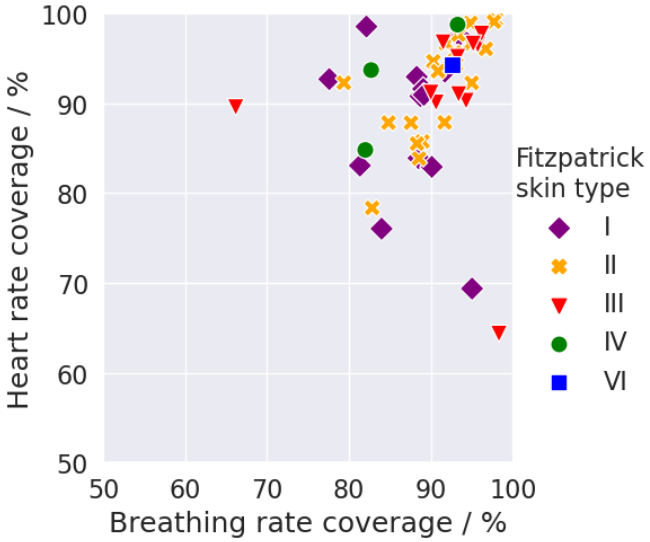



Figure 3.4. Coverage of video-derived vital-sign estimates.

Table 3.3. Coverage statistics for video-derived vital signs.

| | Mean ± Std | Min | Max |
|---|---|---|---|
| BR coverage (%) | 91.0±7.5 | 66.2 | 98.3 |
| HR coverage (%) | 90.0±6.0 | 64.3 | 99.3 |

### 3.3.5 Vital-sign representation learning

State-of-the-art EEG-based sleep staging performance has been achieved using supervised deep learning methods trained with large datasets, often containing thousands of PSG recordings [163]. These datasets e.g. SHHS [170], required multiple years to collect across several data collection centres.

By training a model on a large, source dataset and fine-tuning on a smaller target dataset it is often possible to achieve greater performance than training directly on the target dataset. This approach has successfully been applied in the context of wearable sleep staging [174]. The authors first train a model to classify sleep on a large ECG dataset before fine-tuning it on a much smaller target

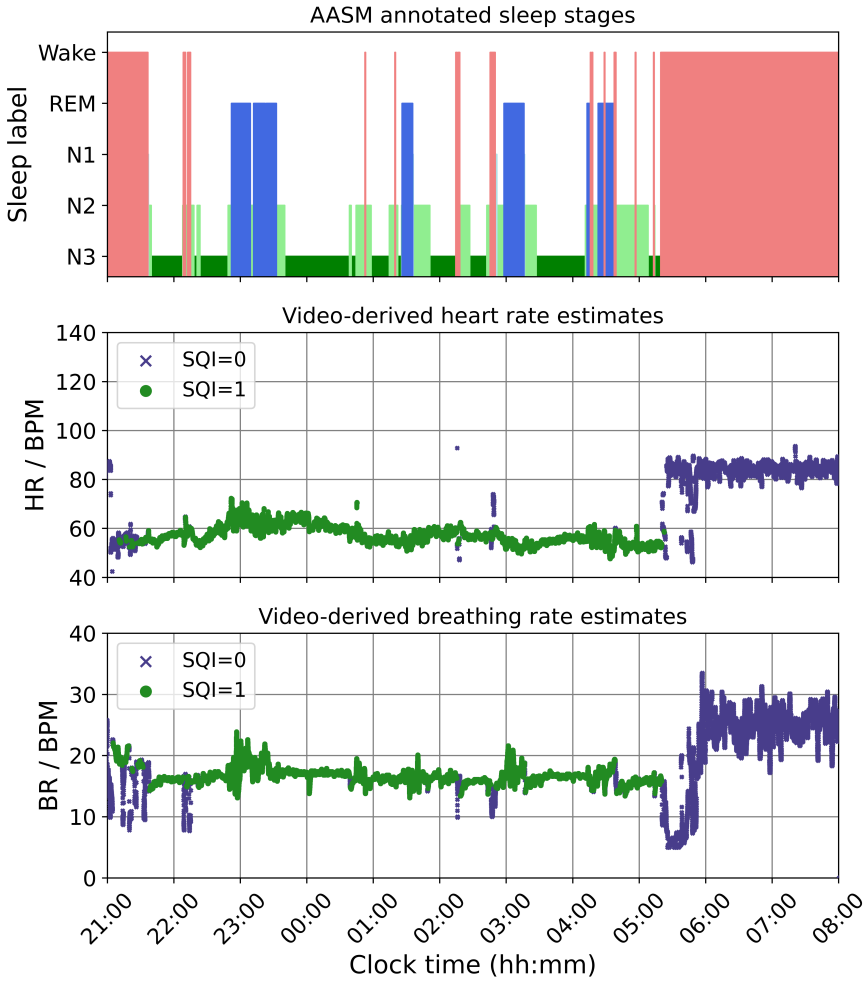



Figure 3.5. Example AASM sleep stages and video-derived heart rate and breathing estimates for a night's recording from the video dataset.

photoplethysmography (PPG) dataset. This is shown to give improved performance over models trained solely on the target PPG dataset.

We use the SHHS dataset [170] to learn representations from heart rate and breathing rate time-series, acquired using wearable sensors, which can then be applied to similar time series obtained from video data. The SHHS dataset is substantially larger than our target video dataset, with polysomnogram recordings and accompanying AASM sleep stage labels from around 6000 participants.

**Training inputs.** Firstly, heart rate (HR) and breathing rate (BR) time-series for the SHHS dataset were estimated from the ECG and RIP signals respectively. An SQI was also calculated for each estimate. More information on our estimation

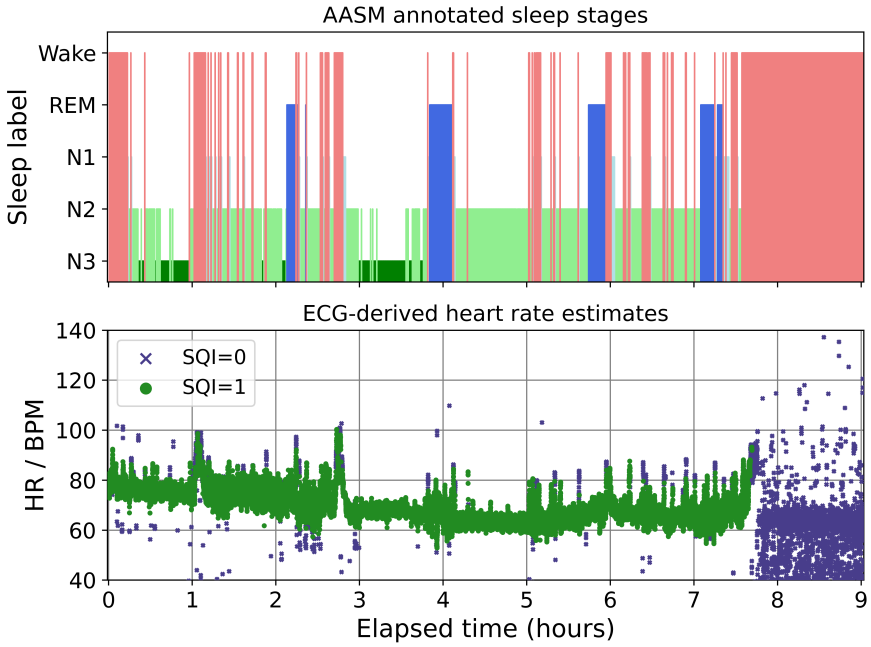



Figure 3.6. Example AASM sleep stages and ECG-derived heart rate estimates and SQI values from an SHHS polysomnogram recording.

process is given in Section 3.A.1. Figure 3.6 shows synchronised AASM sleep stage labels and ECG-derived heart rate estimates for an example recording from the SHHS dataset. We see that there is a long period of low signal quality (SQI=0) at the end of the recording after the participant has woken up. Periods of low signal quality are strongly correlated with wakefulness, since they are often induced by motion artefacts.

**SQI Filtering.** To improve performance, we set low quality sections of the time series to zero before passing them to the model. By zeroing out these sections, we still allow the model to know which periods are of low quality whilst preventing the model from overfitting to signal noise.

**Model architecture.** Figure 3.7 illustrates our neural network architecture. The model uses two input streams to classify each epoch: a high-frequency 5-minute HR/BR window sampled at 1 Hz and a lower frequency 50-minute window sampled at 0.1 Hz, both centred around the target epoch for classification. Each window of data is passed to a 1D ResNet [78] model, which transforms the two-channel input time series into a feature vector. Our ResNet architecture broadly follows the design

of the original ResNet paper, except that 2D convolutions are replaced with their 1D equivalents. The feature vectors are concatenated and passed to a multi-layer perceptron (MLP), which mixes short-term and long-term information from the windows to produce a single feature vector. Finally, a linear classification layer is applied to the feature vector output of the MLP to obtain sleep stage probabilities.

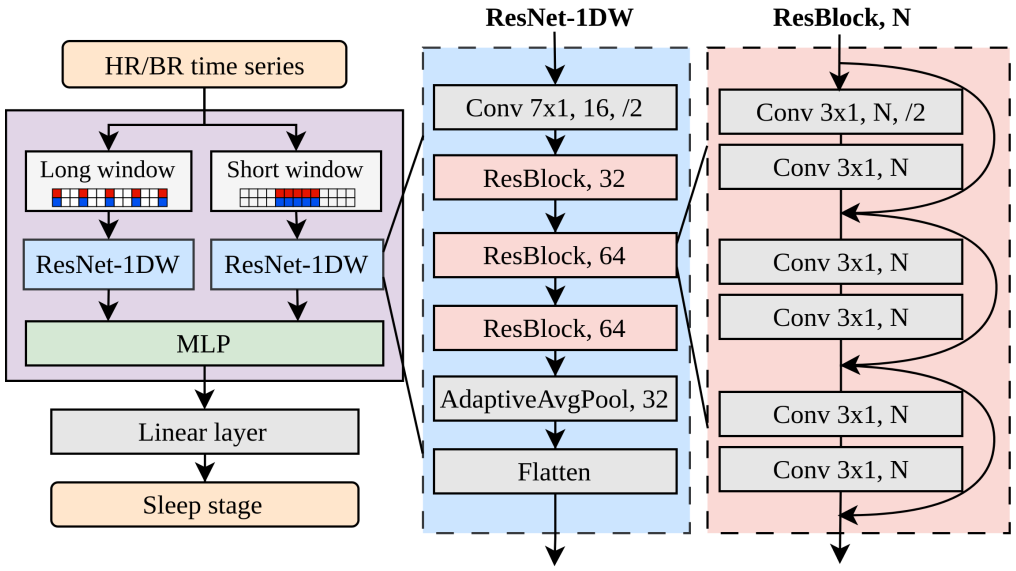



Figure 3.7. Sleep staging model architecture used to learn vital-sign representations from the SHHS dataset. When applying the model to video data, we use the outputs of the MLP (purple box) as input features to an alternative classifier which also incorporates motion information.

**Model training.** The network was trained end-to-end using back-propagation, minimising the cross-entropy loss function against the annotated AASM (5-class) sleep stages. Further model implementation and training details are given in Section 3.A.2. After training, we expect sleep stages to be linearly separable from the output features of the MLP, thus making them effective inputs to any alternative classification model. We use the trained model as a feature extractor by taking the outputs from the MLP layer.

### 3.3.6 Transfer learning

From our video-derived heart rate and breathing rate time series, we use the trained feature extractor to produce 'deep' vital-sign features. These are used

in conjunction with the motion features described in Section 3.3.3 as inputs to a random forest classifier which is trained to classify sleep stages on an epoch-by-epoch basis. A random forest classifier was chosen for its combination of low computational cost and robustness to overfitting.

### 3.3.7 Evaluation

The model was trained and evaluated using k-fold cross-validation, using 10 non-overlapping folds each containing 5 recordings. Performance on each fold was evaluated using a model trained on the remaining 9 folds. Results were then aggregated over all folds.

## 3.4 Results and Discussion

In this section, we report performance in terms of multi-class classification accuracy and Cohen's kappa between model output classifications and annotated sleep stages. A comparison with existing video-based four-class sleep staging algorithms is given in Table 3.4.

Table 3.4. Performance comparison between our deep transfer learning method and prior video-based four-class sleep staging work.

| Method | N[1] | Cohen's $\kappa$ | Accuracy (%) |
|---|---|---|---|
| Nochino *et al.* [153] | 6 | 0.19±0.04 | 40.5±2.2 |
| van Meulen *et al.* [226] | 46 | 0.49±0.13 | 67.9±8.7 |
| This work | 50 | **0.61**±0.15 | **73.4**±9.6 |

[1]Study population size. Mean (± std) across recordings.

Figure 3.8 shows the total test confusion matrix between scorer and model labels summed over all recordings.

**Performance with age.** Figure 3.9 shows the variation in Cohen's kappa values with age. Here we observe a clear negative correlation between age and the Cohen's kappa statistic, findings which mirror those from prior ECG-based sleep staging work [173]. In older adults, common measures of autonomic function such

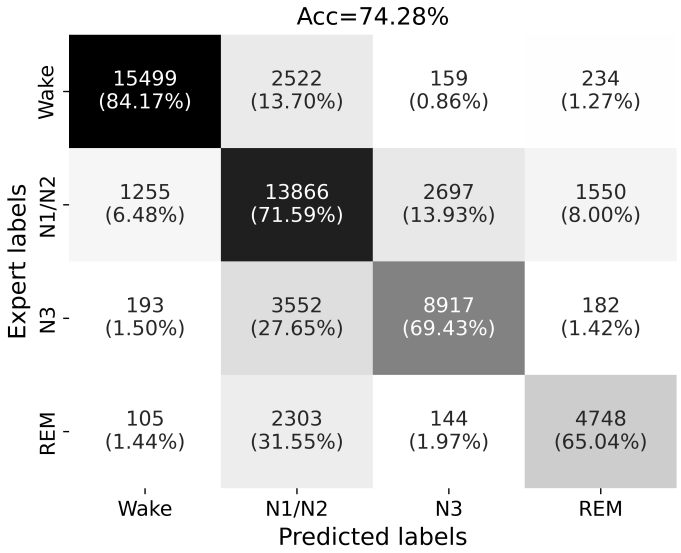



Figure 3.8. Four-class sleep staging confusion matrix across the dataset using our best performing method. Row-wise percentages are given in brackets, with values on the diagonal indicating class-wise sensitivities.

as HRV are known to decrease with age [263], making sleep stages more difficult to distinguish when using cardio-respiratory inputs.

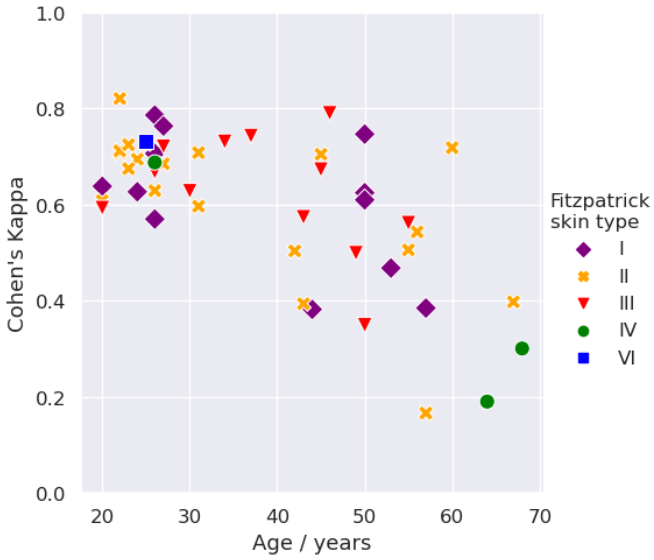



Figure 3.9. Scatter plot of Cohen's $\kappa$ values against age and Fitzpatrick skin type.

**Example hypnogram output.** Figure 3.10 shows an example four-class sleep hypnogram produced by our model. This example corresponds to the median Cohen's kappa obtained across recordings. There is good visual agreement in the

sleep architecture: the model correctly identifies all three REM cycles, sleep onsets and offsets, and most brief arousals during the night. Hypnograms corresponding to the minimum and maximum Cohen's kappa are given in Section 3.A.4.

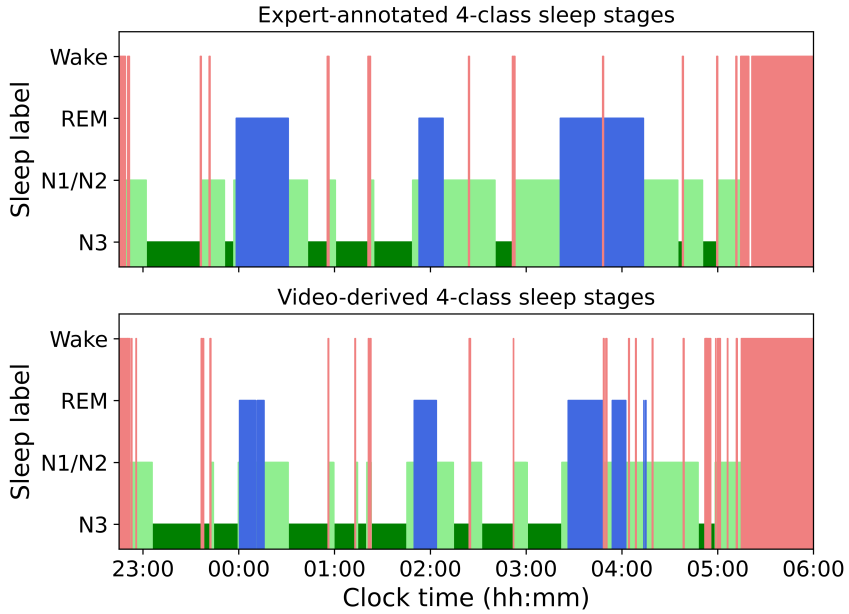



Figure 3.10. Model and scorer four-class hypnograms corresponding to the median Cohen's kappa statistic (0.63).

**Performance with classification strategy.** Table 3.5 indicates the performance of our model using alternative sleep stage classification strategies. The random forest classifier was re-trained for each strategy against the target labels which are derived from the AASM labels accordingly.

Table 3.5. Performance comparison across Sleep–Wake (SW), Wake–NREM–REM (3-class) and Wake–N1/N2–N3–REM (4-class) classification strategies.

| | Cohen's $\kappa$ | Acc. (%) | TPR[1] (%) | TNR[2] (%) |
|---|---|---|---|---|
| SW | 0.72±0.24 | 91.6±6.9 | 94.2±5.3 | 81.0±25.0 |
| 3-class | 0.68±0.18 | 84.9±7.1 | - | - |
| 4-class | 0.61±0.15 | 73.4±9.6 | - | - |

[1]True positive rate i.e. sensitivity. [2]True negative rate i.e. specificity.

Table 3.6. Ablation results showing the effect of adding motion features (M) and of using deep features (DF) extracted from the vital signs compared with a simple feature engineering (FE) baseline.

| Method | Cohen's $\kappa$ | Accuracy (%) |
|---|---|---|
| FE | 0.38±0.20 | 57.6±13.6 |
| FE + M | 0.56±0.17 | 70.2±10.2 |
| DF | 0.50±0.16 | 65.5±10.5 |
| DF + M | 0.61±0.15 | 73.4±9.6 |

Table 3.7. Ablation results showing the effect of SQI filtering.

| Method | Cohen's $\kappa$ | Accuracy (%) |
|---|---|---|
| Raw time-series | 0.59±0.16 | 72.4±10.0 |
| SQI filtering | 0.61±0.15 | 73.4±9.6 |

### 3.4.1 Ablation studies

**Addition of motion and deep features.** In Table 3.6 we assess the impact of including motion features and of using our deep vital-sign features over a simpler baseline of hand-designed features (see Section 3.A.3). We observe that both the deep features and motion features are important to the overall performance.

**SQI filtering.** Table 3.7 indicates the performance improvement obtained from our SQI filtering approach. By zeroing out low quality sections of the time series in both training and testing we achieve an improvement in our results.

## 3.5 Conclusions

In this work, we have introduced a novel transfer learning approach to video-based sleep stage classification which uses non-contact measurements of the heart rate, breathing rate and activity measured using a near-infrared camera. This approach achieves state-of-the-art performance on video-based sleep staging on a healthy study population.

Performance is likely to be further improved through the use of an explicit sequence–sequence architecture such as a Transformer [228], which has led to superior performance in PSG-based sleep staging [166]. Using deep motion features,

such as the intermediate outputs of a sleep pose detection model [144], may also improve classification performance over the relatively simple measures of activity used in this work. In future work, we intend to investigate the performance of our approach across a broader demographic range and in a sleep-disordered population.


# Acknowledgements

This work was supported by the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems [EP/S024050/1] and funded by Oxehealth Ltd. We gratefully acknowledge the National Sleep Research Resource [260] for granting access to the SHHS dataset.


# 3.A Chapter Appendix

## 3.A.1 Vital-sign estimation from contact sensors

**Heart rate:** We estimated the heart rate from the 128 Hz ECG signal available from each PSG recording. First, we applied the Pan-Tompkins algorithm [158] to produce a QRS detection time series. We then applied an FFT to 9-second rolling windows, taking the frequency of the main spectral peak as our estimate. An SQI was calculated for each heart rate estimate given by:

$$\mathrm{SQI}_t^{\mathrm{ECG}} = \begin{cases} 1 & \text{if } |\mathrm{HR}_t^{\mathrm{ECG}} - \mathrm{HR}_t^{\mathrm{PPG}}| \leq 3\,\mathrm{bpm} \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

i.e. our SQI is 1 when there is good agreement between our ECG-derived heart rate, $\mathrm{HR}_t^{\mathrm{ECG}}$, and a PPG-derived heart rate, $\mathrm{HR}_t^{\mathrm{PPG}}$, which is available from the SHHS dataset.

**Breathing rate:** We used a peak counting approach to estimate the breathing rate from both the abdomen and thorax RIP waveforms, using 30-second rolling windows. We then used the thorax-derived estimate as our breathing rate and used the disagreement between the two sources to derive a binary SQI:

$$\mathrm{SQI}_t^{\mathrm{RIP}} = \begin{cases} 1 & \text{if } |\mathrm{BR}_t^{\mathrm{Abd}} - \mathrm{BR}_t^{\mathrm{Thor}}| \leq 3\,\mathrm{brpm} \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

where $\mathrm{BR}_t^{\mathrm{Abd}}$ and $\mathrm{BR}_t^{\mathrm{Thor}}$ are the time-series estimates derived from the abdomen and thorax respectively.

### 3.A.2 Further model implementation details

Each ResNet model produced a flattened 1D feature vector of length 2048. Our MLP therefore had an input dimension of 4096, given by the concatenation of the short- and long-window feature vectors. The MLP used a single hidden layer with a hidden dimension of length 100. The output of the MLP i.e. the 'deep' feature vector had length 32. This was chosen to be small and similar in size to the motion feature vector, to mitigate overfitting and encourage a balanced weighting of the two feature sources in the transfer learning phase.

To improve training stability, we normalised the input time series passed to the model by dividing by the mean and standard deviation of the heart and breathing rates calculated across the entire SHHS dataset. We used the Adam [105] optimiser with default hyper-parameters and employed early stopping after three training epochs without an improvement in validation accuracy.

### 3.A.3 Feature engineering baseline

Our baseline consisted of 6 features extracted from the heart rate and breathing rate time series for each epoch: means over 3-minute rolling windows and standard deviations calculated using 3-minute and 10-minute rolling windows. Each time series was normalised by the 90th percentile value for the entire night and filtered using a difference of Gaussians filter before computing features as done in the wearable sleep staging work of Walch *et al.* [236].

### 3.A.4 Additional model hypnograms

Figure 3.11 shows model and scorer four-class sleep hypnograms for the recording with the highest agreement as measured by Cohen's kappa.

Figure 3.12 shows the hypnograms obtained for the recording with the lowest agreement. In their report, the sleep physiologist noted that there was a 'fuzzy signal

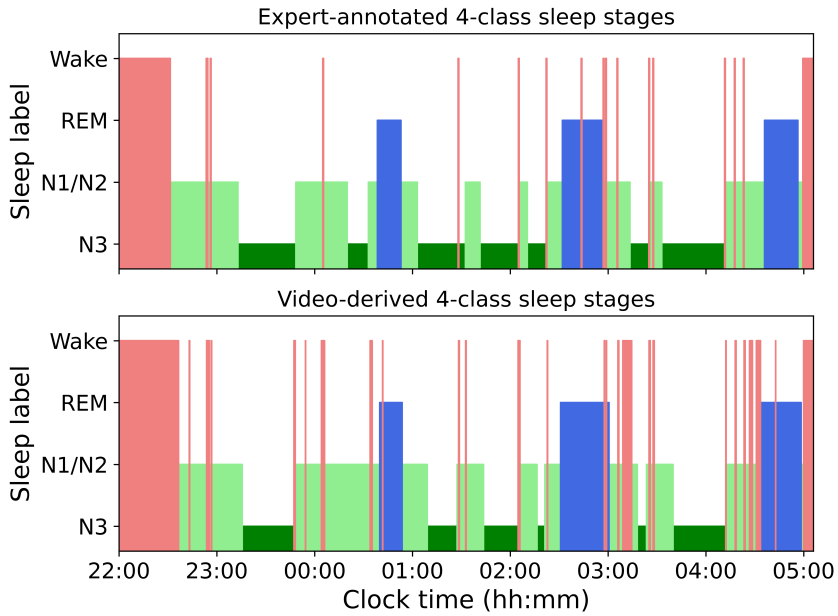



Figure 3.11. Four-class hypnograms corresponding to the highest Cohen's kappa statistic between model and scorer (0.82).

on [many of the EEG channels] throughout, so sleep staging was difficult and not precise'. Measurement factors such as electrode placement and electrical interference are a common source of scorer uncertainty in conventional sleep staging [224].

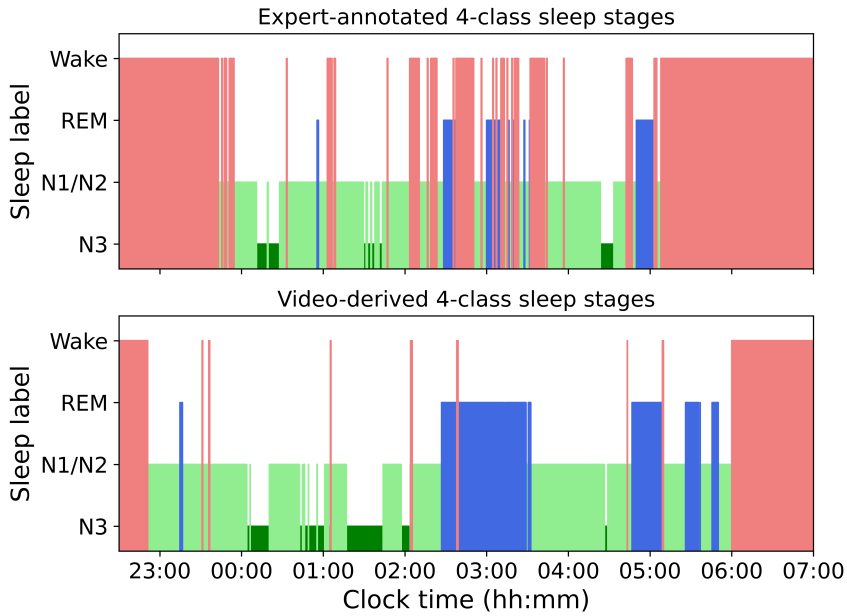



Figure 3.12. Four-class hypnograms corresponding to the lowest Cohen's kappa statistic between model and scorer (0.17).

# 4

# Video-based Sleep Staging from Waveforms

## Preface

This chapter contains the paper “SleepVST: Sleep Staging from Near-Infrared Video Signals using Pre-Trained Transformers". This was presented at the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024 in Seattle where it received a Highlight Paper award. I was the lead author of the paper, devised the proposed methods, conducted all of the experiments, and wrote the initial manuscript. A full Statement of Authorship can be found in Appendix A.

The previous chapter showed that the heart rate, breathing rate and activity can be used to perform video-based sleep staging. This chapter now takes a step further towards clinically useful accuracy. Video-derived cardio-respiratory waveforms are directly used as inputs to a deep learning model, rather than derived heart and breathing rates. This mitigates the loss of important high-frequency signal information such as inter-beat and inter-breath intervals. A novel architecture based on the Transformer [228] is also introduced. These contributions are shown to result in significant improvements in sleep staging accuracy over the methods of the previous chapter.

## Contents

**4.1 Introduction . . . . . 58**
**4.2 Background and Related Work . . . . . 60**
**4.3 Datasets and Preprocessing . . . . . 62**
4.3.1 Contact Sensor Data . . . . . 62
4.3.2 Video Data . . . . . 63
**4.4 Methods . . . . . 64**
4.4.1 SleepVST Architecture . . . . . 64
4.4.2 Motion features from Video . . . . . 66
4.4.3 Transferring SleepVST to Video . . . . . 68
**4.5 Experimental Results . . . . . 69**
4.5.1 Pre-training on Contact Sensors . . . . . 69
4.5.2 Video-based Sleep Staging . . . . . 70
4.5.3 Ablation Studies . . . . . 72
**4.6 Conclusions . . . . . 75**
**4.A Chapter Appendix . . . . . 76**
4.A.1 Broader Impact Statement . . . . . 76
4.A.2 Oxford Sleep Volunteers Dataset . . . . . 76
4.A.3 Additional SleepVST Results . . . . . 77
4.A.4 SleepVST Architecture Details . . . . . 81
4.A.5 Transformer Output Tiling . . . . . 81
4.A.6 Optical Flow Estimation . . . . . 83
4.A.7 Cohen's Kappa Calculation . . . . . 83
4.A.8 SHHS Test Set . . . . . 83

# 4.1 Introduction

Accurate sleep monitoring is critical for the diagnosis of sleep disorders and the discovery of novel treatments and biomarkers. Poor sleep is broadly linked with a number of health conditions, including cardiovascular diseases such as diabetes and hypertension [74]. Additionally, there are links between specific sleep stages and neurodegenerative conditions, such as a decrease in non-rapid-eye-movement stage 3 (N3) sleep with Alzheimer's disease [121].

Overnight video polysomnography (vPSG), the "gold-standard" [207] technique for sleep monitoring, requires the use of a large number of contact sensors including electrodes placed on the scalp (EEG), near the eyes (EOG), and under the chin (EMG) of the patient. Human experts (polysomnographers) must then review the

recorded data, classifying the patient's sleep into five stages at 30-second intervals (epochs) and annotating other important events such as leg movements and apnoeas. This manual process often takes multiple hours to complete.

Experts primarily rely on characteristic patterns in brain activity measured from the EEG to classify sleep stages [94]. However, sleep stage information is also encoded in measures of autonomic activity, including cardiac [195] and respiratory signals [91], and body movements [246]. This has led to the development of methods for automatically classifying sleep stages from sensors that measure these parameters, such as smartwatches [236].

Prior work has shown that these physiological parameters can also be measured using video cameras [80, 240, 241], leading to the development of methods for classifying sleep stages entirely from video input [37, 153, 226]. Camera-based methods have particular promise as part of a multi-purpose sleep monitoring system. For example, within elderly care settings, in addition to classifying sleep stages, they could also be used to detect specific sleep movement disorders [80], which are linked with conditions such as Parkinson's, and which are typically distinguished via manual video review [206]. They could also be used to detect falls [53], one of the leading causes of injury and mortality amongst elderly individuals.

State-of-the-art video-based sleep staging [37] has used the heart rate (HR) and breathing rate (BR) derived from video to perform sleep staging. However, these *rates* are typically measured as averages over time, losing higher frequency inter-pulse and inter-breath interval information. EEG- [165] and wearable-based [110] sleep staging methods have both been improved by using raw waveforms as inputs, rather than derived quantities e.g. signal features [159, 179].

In this paper, we introduce SleepVST[1], a transformer model for sleep stage classification from cardio-respiratory waveforms. First, we pre-train the model on waveforms derived from contact sensors (Fig. 4.1a), after which it outperforms existing methods for cardio-respiratory sleep staging on two widely used public datasets, highlighting the effectiveness of our transformer architecture.

[1]Sleep Video-Signal Transformer.

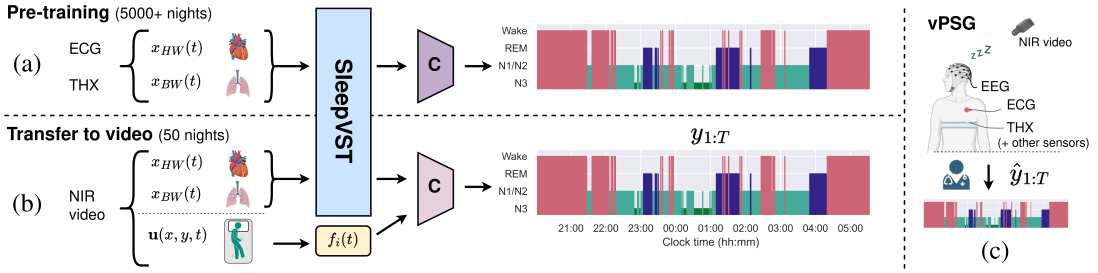



Figure 4.1. **Sleep staging from near-infrared video signals using SleepVST.** (a) We first pre-train the model end-to-end on cardiac (heart) and respiratory (breathing) waveforms $x_{HW}(t)$ and $x_{BW}(t)$ derived from the electrocardiogram (ECG) and a thoracic respiratory belt (THX) respectively. (b) After pre-training, we use the model as a frozen feature extractor, applying it to cardio-respiratory waveforms derived from near-infrared (NIR) video to generate sequences of features. When transferring to video data, we additionally use a set of motion features $f_i(t)$ derived from an optical flow field $\boldsymbol{u}(x, y, t)$ as inputs to the classifier **C**. This approach allows us to utilise much larger contact-sensor datasets to train SleepVST, whilst also enabling the incorporation of motion information when transferring to video data. (c) We evaluate SleepVST using overnight video polysomnography (vPSG) studies, comparing expert-labelled sleep stage sequences $\hat{y}_{1:T}$ against those generated entirely from near-infrared video $y_{1:T}$ using our method.

After pre-training, we then show that SleepVST can be transferred to cardiac and respiratory waveforms measured entirely from near-infrared video, where our method additionally enables us to incorporate motion features derived from the camera data (Fig. 4.1b). Using this novel combination of inputs available from a near-infrared camera, we obtain state-of-the-art results for video-based sleep staging. Through a series of ablation studies we demonstrate the effectiveness of our approach. Our work further highlights the potential of camera-based sleep monitoring for important clinical applications.

## 4.2 Background and Related Work

**Automated sleep staging.** To reduce the manual effort of sleep assessment, deep learning methods have been developed for the automated classification of sleep stages from polysomnography (PSG) recordings, achieving near-human accuracy [163]. Accuracy is commonly measured using Cohen's kappa statistic ($\kappa$, [46]), which measures inter-rater agreement between two sources. Expert annotation of sleep stages is a subjective process, making 100% agreement unachievable [224]. Prior work

has shown that sleep scorers agree around 80% of the time ($\kappa$ = 0.71-0.81, [52, 122]) when classifying sleep stages according to American Academy of Sleep Medicine (AASM, [94]) scoring rules. These rules divide sleep into five classes: Wake, N1 (light), N2 (intermediate), N3 (deep) and rapid-eye-movement (REM) sleep. However, due to low inter-scorer agreement in identifying N1 [52], prior work has often treated sleep classification as a four-class problem, by merging N1 and N2 into a single class [110, 203].

Given the drawbacks of conventional polysomnography, prior work has investigated the potential of automatic sleep staging from a reduced number of contact sensors, including the electrocardiogram ([203], $\kappa$ = 0.66), and the photoplethysmogram ([174], $\kappa$ = 0.65).

**Camera-based physiological measurement.** Respiratory measurement is possible from video by measuring the induced movements of the chest and abdomen [240]. This has been achieved through methods such as frame-differencing [16, 212], using dense optical flow magnitudes [146, 147] and by tracking the movement of sparse optical flow vectors [125] across video frames.

Camera-based measurement of the pulse is possible using methods which measure specific cardiac phenomena [141]. These include remote photoplethysmography (rPPG, [168, 241]), which measures skin colour changes caused by variations in blood volume over the cardiac cycle, and ballistocardiography [19], which measures head movements induced as blood is pumped through the carotid arteries. More recently, deep learning methods have been developed which learn to extract a pulse signal directly from input video frames [43, 129, 257] rather than targeting specific underlying cardiac phenomena, including unsupervised approaches [202, 253].

Camera-based vital-sign measurement has been validated within clinical settings [217, 232], during overnight polysomnography recordings [223, 238], and has led to products with regulatory clearance for camera-based measurement of the heart and breathing rate, including the Gili Pro Biosensor [220] and Oxehealth Vital Signs software [219].

**Camera-based sleep monitoring.** Prior work has investigated the binary problem of distinguishing sleep from wake using video cameras, using measures of activity derived from the input frames [49, 147, 191]. Whilst activity information alone can be effective in distinguishing sleep from wake, cardio-respiratory signals carry important additional information for classifying specific sleep stages [91, 195]. The current state-of-the-art approach for video-based sleep staging ([37], $\kappa = 0.61$) combines multiple modalities derived from video: the heart rate, breathing rate and activity information, to perform sleep stage classification. This approach outperforms prior methods which had used only a single video-derived modality: motion [100, 153] or the cardiac pulse [226].

## 4.3 Datasets and Preprocessing

### 4.3.1 Contact Sensor Data

To train the SleepVST model, we used the SHHS [170, 260] and MESA [44] contact sensor datasets. These datasets contain PSG recordings and sleep stages annotated according to AASM scoring rules. The cardio-respiratory signals available in each recording include the electrocardiogram (ECG), which measures electrical activity of the heart, and respiratory signals, measured using respiratory bands placed around the abdomen and thorax.

From the available signals, we seek to produce generic cardiac and respiratory waveforms to train the SleepVST model, preventing the model from learning to use information which may be present in signals derived from contact sensors but not present in those derived from video. This is to mitigate the problem of domain shift [136, 194], where the performance of a model is significantly reduced when applied to a different data distribution.

To produce a cardiac (heart) waveform, we first applied the same filtering steps used by the Pan-Tompkins algorithm [158] to the ECG signal. We resampled the output from the moving-window integration step of the algorithm to 10 Hz, then applied a Butterworth bandpass filter with a passband of 0.66–2.8 Hz (40–168 beats per minute or BPM) and Gaussian smoothing. To produce a respiratory

(breathing) waveform, we used the thoracic respiratory signal as the input source, as previously used in [17]. This was downsampled to 5 Hz and smoothed using a median filter with a kernel of length 5, to remove artefacts.

### 4.3.2 Video Data

We used the dataset previously introduced in [37] for our video-based sleep staging experiments. This dataset consists of overnight vPSG recordings from 50 volunteers, with sleep stages annotated according to the AASM guidelines at 30-second intervals (epochs) by a sleep physiologist. Throughout this work, we refer to this as the Oxford Sleep Volunteers (OSV) dataset. Each recording includes 850 nm near-infrared (NIR) video data captured at 20 FPS. An example camera view can be seen in Figure 4.5. More information on the dataset is provided in Section 4.A.2.

To derive cardiac and respiratory waveforms from the video data, we used components of FDA-cleared Oxehealth Vital Signs software [219]. For each overnight video recording, we used the software to obtain cardiac and respiratory waveforms sampled at 10 Hz and 5 Hz respectively. We then applied the same filtering steps applied to the contact sensor waveforms: Butterworth filtering and Gaussian smoothing to the heart waveforms, and median filtering to the breathing waveforms.

Figure 4.2 shows examples of processed cardio-respiratory waveforms measured from contact sensors and video for a short section of a recording from the OSV dataset. These have additionally been normalised to zero mean and unit variance.

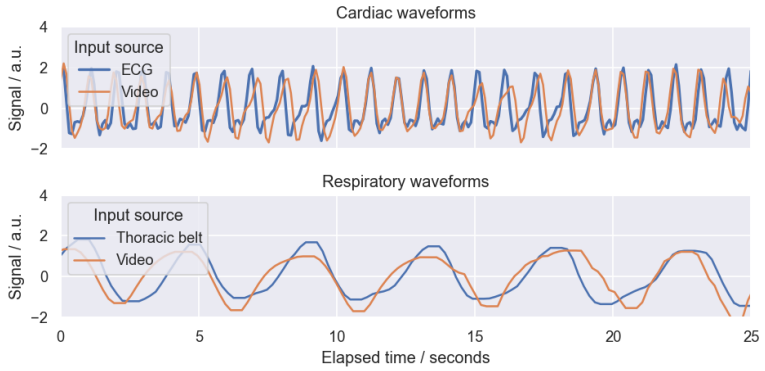



Figure 4.2. Example (normalized) cardiac and respiratory waveforms, $x_{HW}(t)$ and $x_{BW}(t)$, derived from contact sensors (blue) and video (orange) from the OSV dataset.

## 4.4 Methods

### 4.4.1 SleepVST Architecture

Figure 4.3 illustrates the SleepVST model architecture, which turns cardiac and respiratory waveforms into sequences of feature vectors using a transformer encoder. These feature vectors are then used to classify each 30-second sleep epoch. During pre-training, we used a linear layer for classification. When transferring to video data, we trained a new classifier that additionally used motion features derived from video as inputs, discussed further in Section 4.4.2.

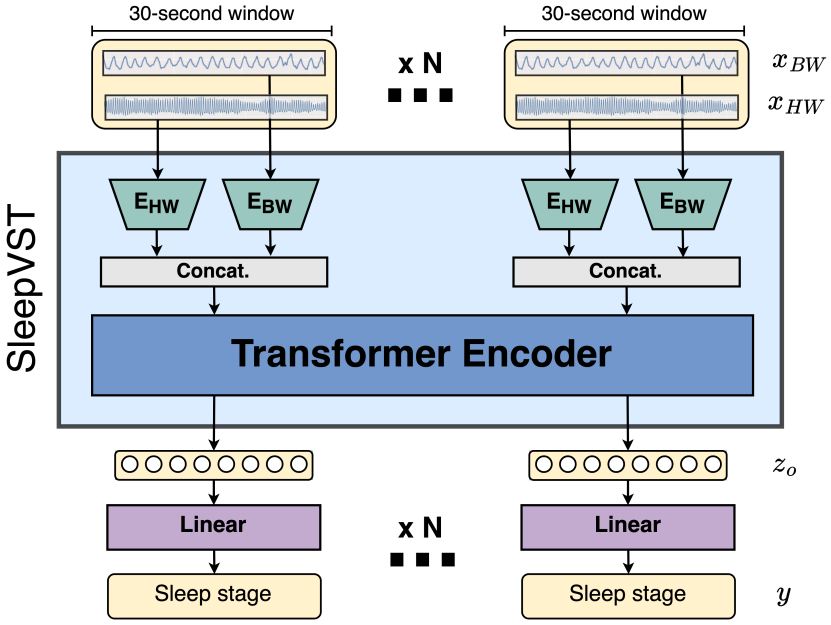



Figure 4.3. **SleepVST architecture.** Each 30-second window of heart ($x_{HW}$) and breathing waveforms ($x_{BW}$) is passed to a **patch encoder**, which turns them into patch-level features. These features are concatenated and passed to a **transformer encoder**. During pre-training, a **linear layer** turns the output feature sequence $z_o$ of length N from SleepVST into sleep stage classifications.

**Patchification.** The continuous heart and breathing waveforms (HW and BW) are divided into N windows i.e. patches, drawing on ideas from PatchTST [151]. We used non-overlapping 30-second windows since this is the interval at which sleep stages are classified according to AASM guidelines. We denote the patchified inputs as $\boldsymbol{x}_{HW} \in \mathbb{R}^{300\times N}$ and $\boldsymbol{x}_{BW} \in \mathbb{R}^{150\times N}$ respectively, where N is the sequence length.

**Encodings.** Each waveform patch is normalised to zero mean and unit variance before being passed to encoding layers $E_{HW}$ and $E_{BW}$, which transform them into feature vectors $\boldsymbol{z}_{HW} \in \mathbb{R}^{D_{HW} \times N}$ and $\boldsymbol{z}_{BW} \in \mathbb{R}^{D_{BW} \times N}$. The feature vectors from each patch are then concatenated, producing a sequence of feature vectors $\boldsymbol{z}_i \in \mathbb{R}^{(D_{HW}+D_{BW}) \times N}$ which are passed to a transformer encoder.

We used identical, shallow 1D ResNet-style [78] models for $E_{HW}$ and $E_{BW}$, illustrated in Figure 4.4. This convolutional encoder design [15] compresses the waveform patches into lower-dimensional feature vectors. This allowed us to keep $D_{HW}$ and $D_{BW}$ small, reducing the computational and memory complexity of the downstream transformer and the likelihood of over-fitting, particularly when transferring the model to the smaller video dataset.

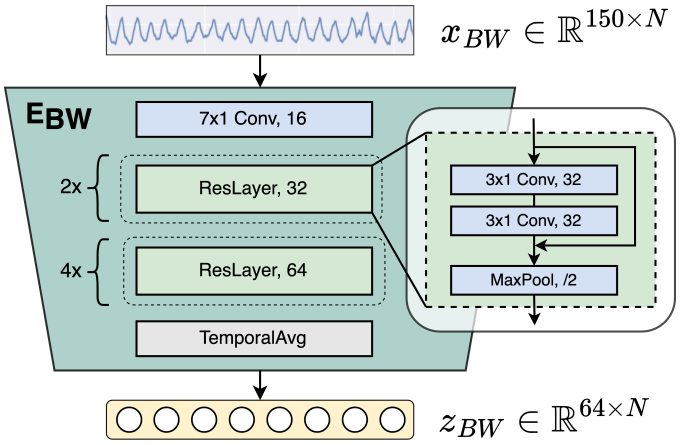



Figure 4.4. **Waveform encoder design.** Using a series of convolutional layers, the encoder turns sequences of signal patches into sequences of lower-dimensional feature vectors.

This encoder design additionally provides a high degree of flexibility at the input level, such as the ability to use input signals of differing frequencies. In later ablation studies, this enables us to compare the performance of the SleepVST model with simpler variants which use 30-second patches of 1 Hz heart rate $\boldsymbol{x}_{HR} \in \mathbb{R}^{30 \times N}$ and/or breathing rate $\boldsymbol{x}_{BR} \in \mathbb{R}^{30 \times N}$ time-series as inputs, rather than underlying waveforms. When using either of these as inputs, we used linear projections [59, 151] for the input encoders, because of the lower dimensionality of

$\boldsymbol{x}_{HR}$ and $\boldsymbol{x}_{BR}$. Throughout our experiments we use the same feature dimension i.e. $D_{HW} = D_{BW} = D_{HR} = D_{BR} = 64$.

**Transformer layer.** We used a Transformer encoder [228] with sinusoidal position encodings to turn the sequence of input features $\boldsymbol{z}_i$ into an output sequence of features $\boldsymbol{z}_o$ with an identical feature dimension. Normal sleep exhibits long-range structures, known as sleep cycles [161], which typically last around 90–120 minutes and typically occur four to five times per night. From its superior performance in many sequence modelling tasks, including EEG-based sleep staging [166], we expect that the inductive biases of the transformer model (i.e. the attention mechanism) can enable it to learn to use relative position within a sleep cycle to classify sleep stages.

**Classification.** When training and testing on contact sensor data, we used a linear classification layer $\boldsymbol{W}_C \in \mathbb{R}^{C\times(D_{HW}+D_{BW})}$ to transform the sequence of output feature vectors $\boldsymbol{z}_o$ into a sequence of sleep stage probabilities $\boldsymbol{y} \in \mathbb{R}^{C\times N}$, where the number of classes $C$ depends on the sleep classification strategy used. In line with prior work, we classified sleep stages into four classes during pre-training: Wake–N1/N2–N3–REM, i.e. $C = 4$.

**Model training and inference.** In all experiments, we used a sequence length of N = 240 windows i.e. two hours. During training, we created batches by sampling sequences of length N with a step of 5 minutes from within each recording. The model was trained to minimise the multi-class cross-entropy loss using the AdamW [134] optimiser with a learning rate of 0.0005 and weight decay of 0.001. Other hyper-parameters, such as the no. of encoder layers, are detailed in Section 4.A.4. To apply the model to arbitrary length input sequences during evaluation, we re-applied it at regular intervals. More details on this process are provided in Section 4.A.5.

### 4.4.2 Motion features from Video

Following prior work [153], we seek to produce motion features from video which encode for information which is predictive of the sleep stage, such as the rate and distribution of body movements [246]. To do so, we define a set of parameterised

time-series features that collectively quantify motion over a range of spatial and temporal scales.

**Optical flow estimation.** We first estimated a 2D optical flow field $\boldsymbol{u}(t, x, y) \in \mathbb{R}^2$ over the bed region. Inspired by sleep pose detection work [144], we applied a homography transformation [77] to the individual video frames before estimating the flow field, to achieve camera viewpoint invariance. We additionally divide each frame into approximate head (H), body (B) and outer bed (O) regions R for further processing. This is illustrated in Figure 4.5.

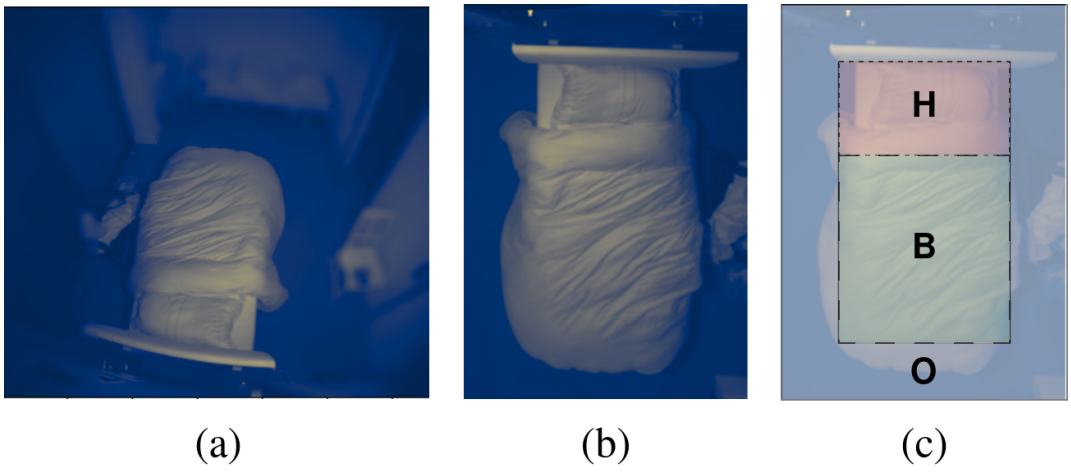



Figure 4.5. **Example processing of an NIR video frame from the OSV dataset.** (a) Real viewpoint. (b) Virtual viewpoint. (c) Head (H), body (B), and outer (O) bed regions. The distance from the camera to the head region is around 1.5 m.

**Feature definitions.** Given a region R of the optical flow field, we define the following parametrised signals:

$$v(t; R) = \max_{(x,y)\in R} |\boldsymbol{u}(t, x, y)| \tag{4.1}$$

$$s(t; R) = \frac{1}{|R|} \sum_{(x,y)\in R} |\boldsymbol{u}(t, x, y)| \tag{4.2}$$

i.e. the maximum (4.1) and the average (4.2) of the optical flow magnitude in the region R at time $t$. Using these signals, we define two time-series features that

quantify motion within a time window $\Delta$ up to time $t$:

$$f_1(t; R, \Delta) = \sum_{t-\Delta}^{t} v(t; R) \tag{4.3}$$

$$f_2(t; R, \Delta) = \sum_{t-\Delta}^{t} s(t; R) \tag{4.4}$$

Plus two additional features which measure the time elapsed since either signal went above a threshold $\delta$:

$$\begin{aligned} f_3(t; R, \delta) &= \sum_{t-\tau}^{t} 1 \\ \text{where} \quad \tau &= \max_{t' \leq t} t' \\ \text{s.t.} \quad & v(t'; R) > \delta \end{aligned} \tag{4.5}$$

$$\begin{aligned} f_4(t; R, \delta) &= \sum_{t-\tau}^{t} 1 \\ \text{where} \quad \tau &= \max_{t' \leq t} t' \\ \text{s.t.} \quad & s(t'; R) > \delta \end{aligned} \tag{4.6}$$

We calculated each feature for all possible combinations of parameters $\Delta \in \{30, 300\}$ seconds, $\delta \in \{0.01, 0.1, 1\}$ and $R \in \{H, B, O\}$. This resulted in (2+2+3+3) time-series features per region, i.e. 30 unique time-series features. Additionally, we provided the video-based classifier (see Section 4.4.3) with time-shifted features $f_i(t + T)$ for $T \in \{\text{-}90, 0, +90\}$ seconds. This gave a total of 90 motion features for each sleep epoch.

### 4.4.3 Transferring SleepVST to Video

To transfer the SleepVST model to video using the OSV dataset, we removed the linear classification layer learnt during pre-training and used the model as a frozen feature extractor, applying it to sequences of cardiac and respiratory waveforms measured from video to produce sequences of output features $\boldsymbol{z}_o$. We then trained a new classifier that used these features, plus the motion features derived in the previous section, to classify individual sleep epochs. This allowed us to leverage the representational capacity of the transformer model, by pre-training on the much larger contact-sensor datasets, whilst also enabling us to incorporate motion

information from the video dataset. We followed the design choice of [37] and used a Random Forest [33] for our classifier due to its low computational cost and robustness to overfitting.

## 4.5 Experimental Results

Sleep staging performance has commonly been reported in terms of classification accuracy ($\text{Acc}_T$) and total Cohen's kappa ($\kappa_T$) over all sleep epochs in the dataset, and/or using per-night mean statistics, $\text{Acc}_\mu$ and $\kappa_\mu$, i.e.:

$$\kappa_T = K(\sum_{i=1}^{N} c(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i)) \tag{4.7}$$

$$\kappa_\mu = \frac{1}{N} \sum_{i=1}^{N} K(c(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i)) \tag{4.8}$$

where $\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i \in \mathbb{R}^{C \times L_i}$ are the one-hot encoded model and scorer labels for recording $i$ of length $L_i$ and $c(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i) \in \mathbb{R}^{C \times C}$ is the confusion matrix between them. $K(\cdot)$ is the function which calculates the Cohen's kappa statistic [46] from a confusion matrix and is defined in Section 4.A.7. We report all of these statistics, to make fairer comparisons with prior work, and to aid future comparisons.

### 4.5.1 Pre-training on Contact Sensors

From the SHHS dataset, we randomly selected 500 participants who participated in both visits, to form a test set of 1000 nights of data. This ensured that no participant appeared in both the training/validation and test partitions. Our test set is listed in Section 4.A.8. From the MESA dataset, we used the same test set of 204 nights as Kotzen *et al.* [110]. The remaining nights from both the SHHS and MESA datasets were pooled together and randomly split into training and validation sets using a 75:25 split.

Table 4.1 compares the overall performance of SleepVST with prior methods for cardio-respiratory sleep staging from contact sensors. These have used input modalities such as the photoplethysmogram (PPG) waveform and the ECG-derived heart rate. Using the ECG and thoracic respiratory signals, SleepVST outperforms prior methods for sleep staging from cardio-respiratory signals on both datasets.

Table 4.1. Comparison of four-class cardio-respiratory sleep staging methods using contact sensors.

| | | | | Cohen's $\kappa$ | | Accuracy / % | |
|---|---|---|---|---|---|---|---|
| Dataset | N | Method | Modalities | $\kappa_\mu$ | $\kappa_T$ | $\text{Acc}_\mu$ | $\text{Acc}_T$ |
| SHHS [170] | 296 | Bakker et al. [17] | HR+Thor*+Air† | - | 0.64 | - | 76.7 |
| | 800 | Sridhar et al. [203] | HR (ECG) | 0.65 | 0.66 | 77.3 | 77.0 |
| | 1000 | SleepVST | ECG+Thor* | **0.73** | **0.75** | **82.8** | **83.0** |
| MESA [44] | 296 | Bakker et al. [17] | HR+Thor*+Air† | - | 0.68 | - | 79.8 |
| | 194 | Sridhar et al. [203] | HR (ECG) | - | 0.69 | - | 80.0 |
| | 204 | Kotzen et al. [110] | PPG | - | 0.73‡ | - | 82.6‡ |
| | 204 | SleepVST | ECG+Thor* | **0.76** | **0.77** | **85.1** | **85.2** |

*Thoracic respiratory effort. †Nasal airflow. ‡From reported confusion matrices.

### 4.5.2 Video-based Sleep Staging

In this section, we report the performance of the SleepVST model after transfer to the OSV video dataset. We follow the evaluation procedure of [37], training and evaluating on the 50 nights of data using 10-fold cross-validation with non-overlapping folds each containing five nights of data.

In Table 4.2, we see that our method achieves significantly better performance in terms of both accuracy and Cohen's $\kappa$ statistic when compared with prior video-based methods evaluated on similar study populations.

Table 4.2. Comparison of four-class sleep staging methods using video cameras.

| | | | | Cohen's $\kappa$ | | Accuracy / % | |
|---|---|---|---|---|---|---|---|
| Dataset | N | Method | Modalities** | $\kappa_\mu$ | $\kappa_T$ | $\text{Acc}_\mu$ | $\text{Acc}_T$ |
| Healthy infants [100] | 8 | Kamon et al. [100] | Motion | 0.26 | - | 48.0 | - |
| Healthy adults [153] | 6 | Nochino et al. [153] | Motion | 0.19 | - | 40.5 | - |
| HealthBed Study [226] | 46 | van Meulen et al. [226] | HW* | 0.49 | 0.49‡ | 67.9 | 67.9‡ |
| Oxford Sleep Volunteers | 50 | Carter et al. [37] | HR+BR+Motion | 0.61 | 0.64‡ | 73.4 | 74.3‡ |
| | 50 | SleepVST | HW*+BW†+Motion | **0.68** | **0.71** | **77.7** | **78.8** |

**Camera-derived. *Cardiac pulse (heart) waveforms. †Respiration (breathing) waveforms. ‡From reported confusion matrices.

Figure 4.6 shows the total test confusion matrix between model and human expert (polysomnographer) classifications on the OSV dataset. We observe particularly high accuracy in distinguishing Wake and REM sleep stages.

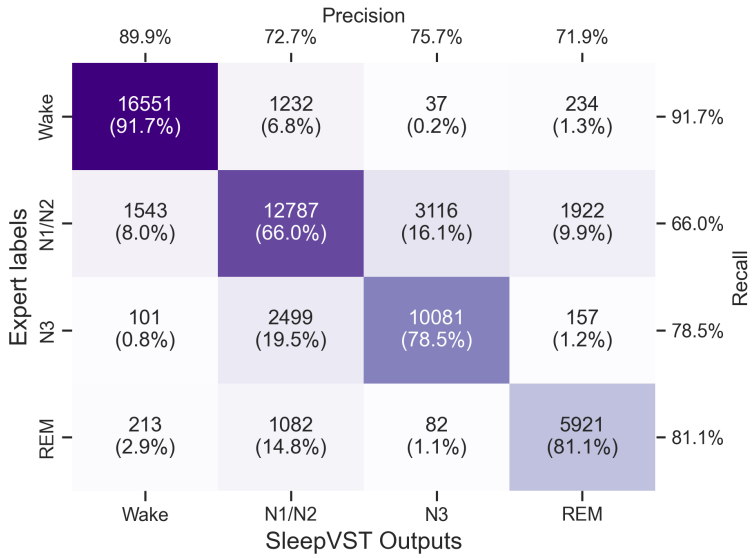



Figure 4.6. Confusion matrix between SleepVST and expert labels on four-class video based sleep staging.

**Example SleepVST output.** Figure 4.7 shows an example hypnogram generated entirely from near-infrared video using our method. This example corresponds to the median Cohen's $\kappa$ between model and expert from the OSV dataset. The model correctly identifies all three REM cycles, periods of deep sleep, and brief awakenings throughout the night. Additional hypnogram examples can be found in Section 4.A.3.

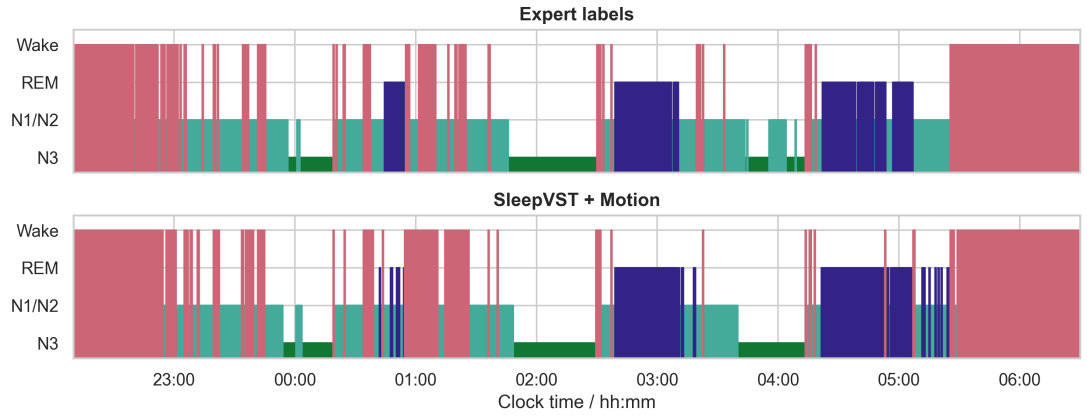



Figure 4.7. **Example four-class sleep hypnograms from the OSV dataset.** (top) Scored by an expert using signals from the vPSG recording. (bottom) Automatically generated from near-infrared video using SleepVST. (Cohen's $\kappa = 0.71$ between model and expert sleep stages.)

**Population performance.** Figure 4.8 shows the distribution of Cohen's $\kappa$ values across participants from the OSV dataset plotted against age, sex and Fitzpatrick skin type[2] [63]. We observe a decrease in performance with age, as observed in prior work [37, 209]. In older adults, measures of cardio-respiratory activity (e.g. heart rate variability) are known to decline with age [263]. This likely reduces the distance between sleep stages in cardio-respiratory input space and thus makes them harder to distinguish. Additional video data from older subjects would likely help to improve this, which is discussed further in Section 4.A.3.

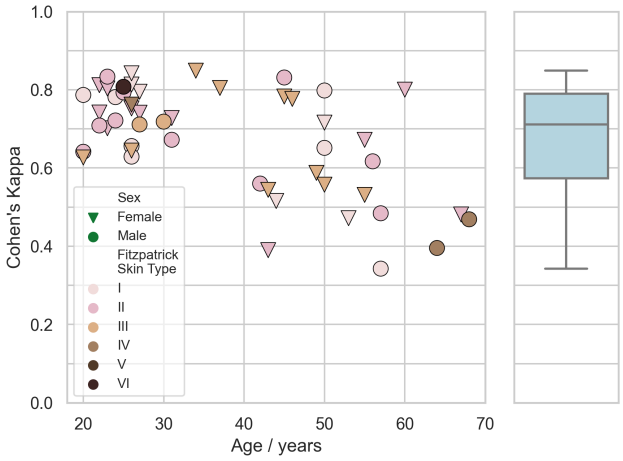



Figure 4.8. Scatter and box plot of Cohen's $\kappa$ against age, sex and Fitzpatrick skin type[2] for individuals from the OSV dataset.

### 4.5.3 Ablation Studies

**Video transfer strategy.** First, we compare the performance of three methods for transferring the SleepVST model to video data:

1. Directly applying the pre-trained SleepVST model to cardiac and respiratory waveforms from video data.

2. Using SleepVST as a feature extractor and:

   (a) training a new classification layer without using motion features.

   (b) training a new classification layer that also uses motion features.

[2]Plotted using representative skin tones from [245].

In Table 4.3, we see that training a new classifier for video data is essential to the effectiveness of our approach. This step likely allows the classifier to ignore parts of the learnt feature space that do not transfer between contact-sensor and video-derived waveforms i.e. sensor-specific features.

Table 4.3. SleepVST video-based sleep staging performance for different model transfer strategies.

| | $\kappa_\mu$ | $\kappa_T$ | $\text{Acc}_\mu$ / % | $\text{Acc}_T$ / % |
|---|---|---|---|---|
| Direct apply | 0.510 | 0.536 | 66.7 | 67.3 |
| Transfer w/o motion | 0.660 | 0.689 | 76.4 | 77.4 |
| Transfer w/ motion | **0.677** | **0.708** | **77.7** | **78.8** |

**Motion features.** Table 4.4 reports the effectiveness of the two main components of our motion feature set: the set of feature definitions $f_i$, and the set of regions used $R$. Additional ablations can be found in Section 4.A.3. Here we define and report an additional metric $\kappa_F$ which is the Cohen's $\kappa$ statistic computed over epochs near transitions to/from Wake ($W$) i.e. where $\hat{y}_i^t \neq \hat{y}_i^{t+\Delta} = W$ or $\hat{y}_i^t = W \neq \hat{y}_i^{t+\Delta}$ for any $\Delta \in [-2, 2]$ epochs. This metric better quantifies accuracy during fragmented sleep, such as in Figure 4.9. As measured by $\kappa_F$, we see that motion features aid classification during fragmented sleep. The ability to accurately measure sleep fragmentation is particularly important since it is linked with daytime sleepiness and decreased cognitive performance [140].

Table 4.4. Ablation study of motion feature parameters.

| | | | Cohen's $\kappa$ | |
|---|---|---|---|---|
| Ablation | Parameter Set | $n_f^*$ | $\kappa_T$ | $\kappa_F$ |
| Feature set | $f_i \in \{\}$ | 0 | 0.689 | 0.397 |
| | $f_i \in \{f_1, f_2\}$ | 36 | 0.697 | 0.467 |
| | $f_i \in \{f_3, f_4\}$ | 54 | 0.704 | 0.465 |
| | $f_i \in \{f_1, f_2, f_3, f_4\}$ | 90 | **0.708** | **0.491** |
| Region(s) | $R \in \{H \cup B\}$ | 30 | 0.702 | 0.461 |
| | $R \in \{H \cup B, O\}$ | 60 | 0.704 | 0.468 |
| | $R \in \{H, B, O\}$ | 90 | **0.708** | **0.491** |

*No. features.

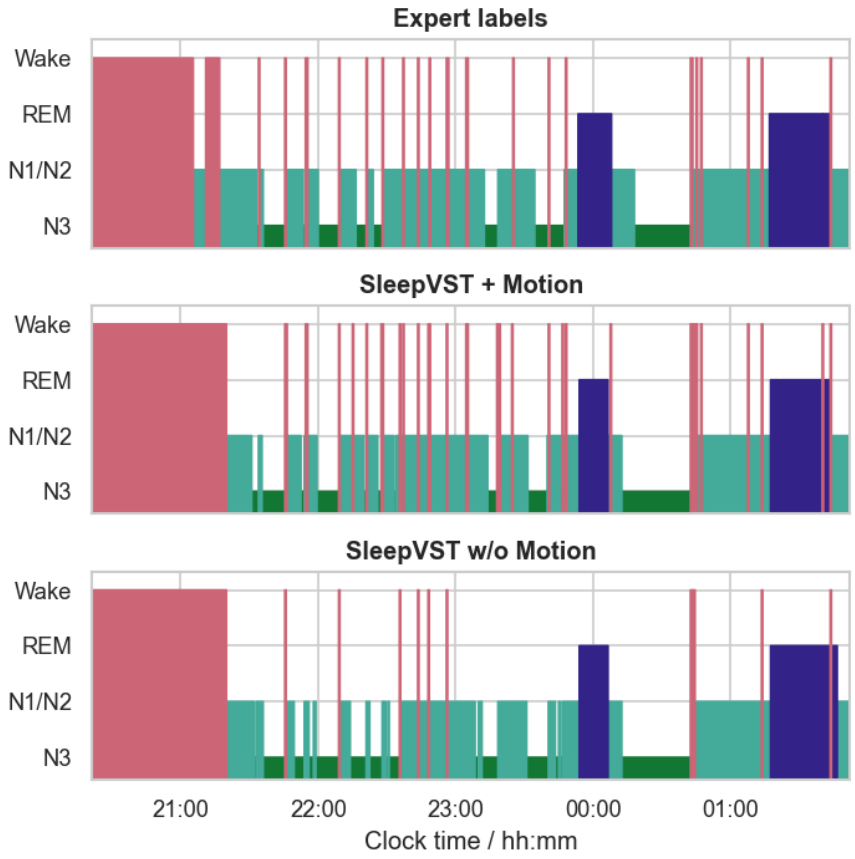



Figure 4.9. **Example four-class hypnograms during fragmented sleep from the OSV dataset.** The use of motion features improves accuracy during short awakenings between 22.00 and 00.00.

**Cardio-respiratory inputs.** In Table 4.5, we compare the performance of the SleepVST model with variants which use heart rate $\boldsymbol{x}_{HR}$ or breathing rate $\boldsymbol{x}_{BR}$ as inputs instead of the underlying waveforms, as described in Section 4.4.1. We see that the use of waveforms, rather than derived heart and breathing rates, leads to improvements in sleep staging accuracy across all datasets.

Table 4.5. Comparison of four-class sleep staging performance for different cardio-respiratory input combinations to SleepVST.

| | Cohen's $\kappa_T$ | | |
|---|---|---|---|
| Input signals | SHHS | MESA | OSV |
| $\boldsymbol{x}_{HR}$ , $\boldsymbol{x}_{BR}$ | 0.683 | 0.722 | 0.681 |
| $\boldsymbol{x}_{HW}$, $\boldsymbol{x}_{BR}$ | 0.718 | 0.736 | 0.684 |
| $\boldsymbol{x}_{HR}$ , $\boldsymbol{x}_{BW}$ | 0.723 | 0.758 | 0.702 |
| $\boldsymbol{x}_{HW}$, $\boldsymbol{x}_{BW}$ | **0.749** | **0.765** | **0.708** |

## 4.6 Conclusions

We have introduced SleepVST, a transformer model for sleep stage classification from cardio-respiratory waveforms. Using a transfer learning approach, we have shown how SleepVST can be used to perform sleep stage classification entirely from near-infrared video.

Our results advance the state-of-the-art in video-based sleep staging, narrowing the gap to expert-level performance and taking a significant step towards important clinical applications. More broadly, our work shows that a model pre-trained on a task using cardio-respiratory waveforms from large contact-sensor datasets can be effectively transferred to waveforms measured from video. This approach can benefit wider applications such as the detection of arrhythmias and sleep apnoeas, from cardiac and respiratory waveforms respectively.

Finally, we believe there is important future work in learning more expressive representations of motion during sleep. In addition to improving the performance of video-based sleep staging, this could also enable the detection of sleep movement disorders, such as REM behaviour disorder, removing the existing need for time-consuming, manual video review [206].

# 4.A Chapter Appendix

## 4.A.1 Broader Impact Statement

Video data is a sensitive modality and its use in healthcare must come in conjunction with patient consent and clinical assessment. In sleep medicine, video polysomnography is considered the gold-standard monitoring technique, and the use of video enables accurate diagnoses of specific conditions such as REM behaviour disorder [206]. For wider applications, the algorithms introduced in this work have been designed such that the video data can be processed into intermediate motion and cardio-respiratory signals for further processing offline. This means it is possible for camera data to be processed in real-time without being stored.

Machine learning methods have the potential to make a transformative impact in healthcare. However, they should be appropriately evaluated across diverse populations and in real-world scenarios, to understand their limitations, mitigate potential biases, and reduce clinical risk. In future work, we plan to evaluate the performance of our method across a broader population, including a greater number of older individuals and individuals with darker skin tones, and individuals with diagnosed sleep disorders.

## 4.A.2 Oxford Sleep Volunteers Dataset

Throughout this work, we used the video polysomnography dataset introduced by Carter *et al.* [37] for our video-based sleep staging experiments, which we refer to as the Oxford Sleep Volunteers (OSV) dataset. Figure 4.10 shows the two camera viewpoints and room layouts present in the dataset, along with transformed frames, and approximate head, body and outer bed regions. The homography transformations and regions were manually determined from the known camera parameters and room geometry. For more information on the dataset, including population demographics and room geometries, we refer to the original work [37].

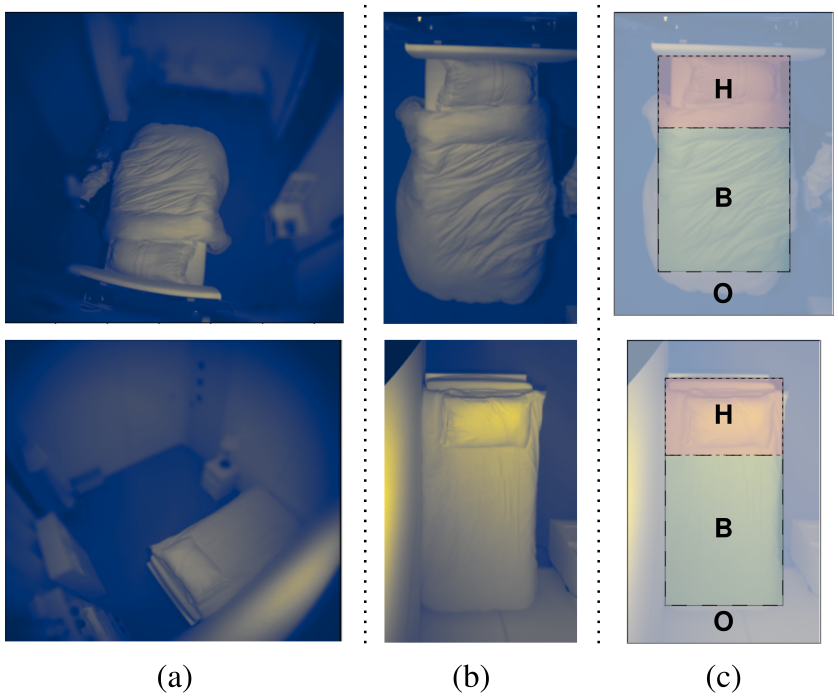



Figure 4.10. **Example processing of near-infrared video frames from the OSV dataset.** (a) Real viewpoints and bed positions for each room. (b) Cropped, virtual viewpoints obtained using homography transformations. (c) Head (H), body (B), and outer (O) bed regions.

### 4.A.3 Additional SleepVST Results

**Cohen's $\kappa$ distribution across datasets.** Figure 4.11 shows the distribution of Cohen's $\kappa$ values with age across all three datasets. Despite being trained on contact sensor waveforms from much older subjects in the SHHS and MESA datasets, SleepVST successfully transfers to video-derived waveforms, nearly reaching parity with performance using contact sensors. This highlights the effectiveness and generality of the learnt feature space. Given the well-known changes in autonomic activity with age [263], pre-training with additional data from younger subjects may further improve transfer learning performance on the existing datasets, as may fine-tuning the entire SleepVST network on video data, rather than freezing the weights and using it as a fixed feature extractor. Conversely, more video data from older participants during the transfer learning phase would likely help to improve the performance for the existing older participants.

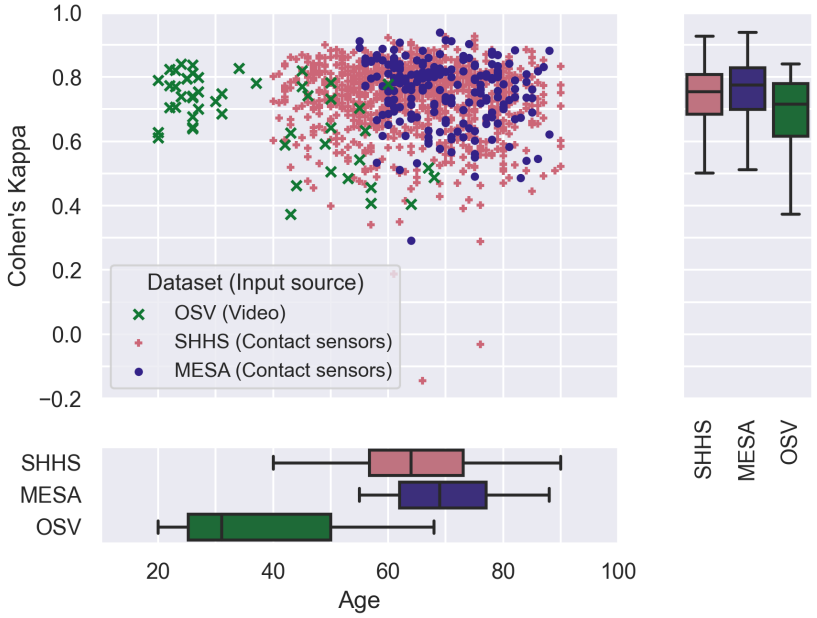
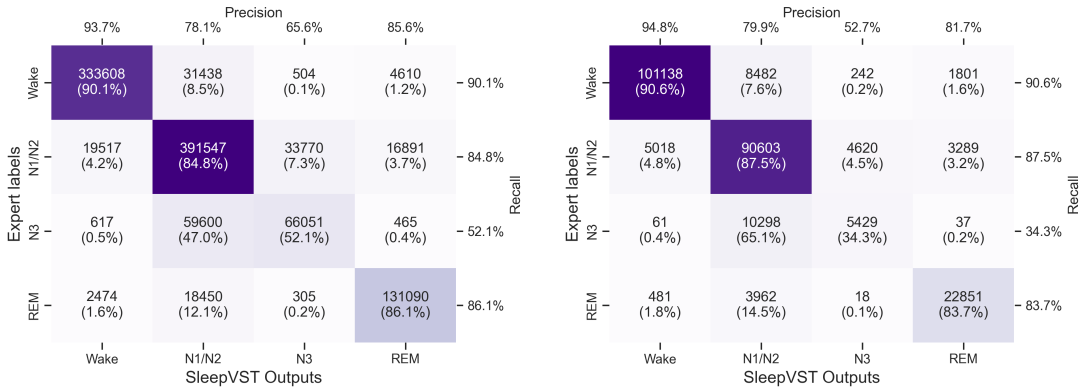



Figure 4.11. Scatter and box plots of Cohen's $\kappa$ and age distributions across the SHHS, MESA and OSV datasets.

**Pre-training confusion matrices.** Figures 4.12a and 4.12b show the four-class sleep staging confusion matrices obtained after pre-training, for the SHHS and MESA test sets respectively. During pre-training, we chose to use an unbalanced cross-entropy loss. Using a class-weighted loss instead would likely help to reduce the observed rate of misclassification of (less frequent) N3 sleep.




(a) SHHS dataset. (b) MESA dataset.

Figure 4.12. SleepVST confusion matrices when applied to the contact sensor test sets after pre-training.

**Varying classification strategy.** In Table 4.6, we report the performance of our method for different sleep stage classification strategies, to aid future comparisons. In each case, we use the same SleepVST model (pre-trained on four-class sleep staging) as a feature extractor, and train a new classifier using video data.

Table 4.6. Video-based sleep staging performance using SleepVST with different classification strategies.

| | $\kappa_\mu$ | $\kappa_T$ | $\text{Acc}_\mu$ / % | $\text{Acc}_T$ / % |
|---|---|---|---|---|
| Sleep–Wake | 0.782±0.197 | 0.857 | 93.3±5.4 | 93.7 |
| W–NREM–REM | 0.760±0.139 | 0.795 | 87.4±6.6 | 87.9 |
| W–N1/N2–N3–REM | 0.677±0.133 | 0.708 | 77.7±8.9 | 78.8 |
| W–N1–N2–N3–REM | 0.646±0.139 | 0.674 | 74.1±10.7 | 75.3 |

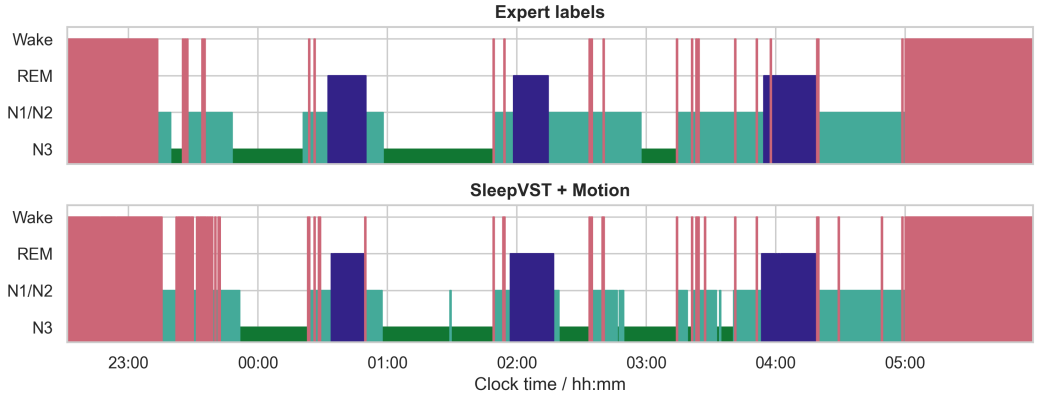
**Additional motion feature ablations.** Table 4.7 shows the effectiveness of the remaining components of our feature set. Higher thresholds ($\delta = 1$) measure the elapsed time since large movements, which are more important for overall agreement $\kappa_T$. Lower thresholds ($\delta = 0.01$) are more sensitive to smaller movements e.g. brief awakenings, leading to greater agreement around fragmented sleep $\kappa_F$.

Table 4.7. Ablation of motion feature parameters $T, \delta$ and $\Delta$.

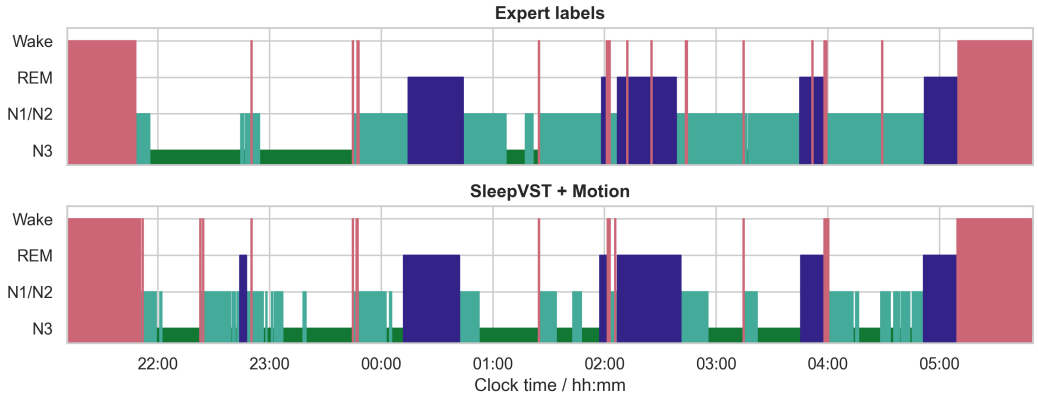
| | | | Cohen's $\kappa$ | |
|---|---|---|---|---|
| Ablation | Parameter Set | $n_f^*$ | $\kappa_T$ | $\kappa_F$ |
| Threshold(s) | $\delta \in \{0.01\}$ | 54 | 0.698 | 0.485 |
| | $\delta \in \{0.1\}$ | 54 | 0.702 | 0.486 |
| | $\delta \in \{1\}$ | 54 | 0.705 | 0.471 |
| | $\delta \in \{0.01, 0.1, 1\}$ | 90 | **0.708** | **0.491** |
| Time shift(s) | $T \in \{0\}$ | 30 | 0.705 | 0.464 |
| | $T \in \{\text{-}90, 0, 90\}$ | 90 | **0.708** | **0.491** |
| Window(s) | $\Delta \in \{30\}$ | 72 | 0.706 | 0.486 |
| | $\Delta \in \{300\}$ | 72 | 0.705 | 0.472 |
| | $\Delta \in \{30, 300\}$ | 90 | **0.708** | **0.491** |

*No. features.

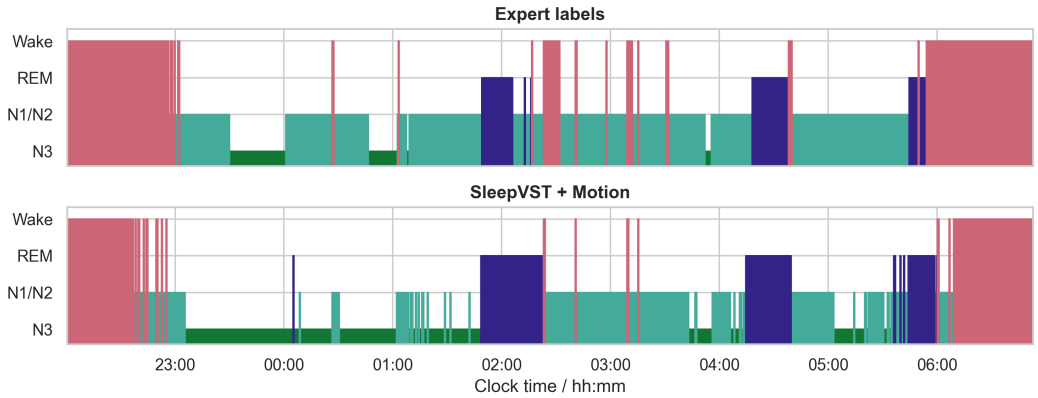
**Additional model hypnograms.** Additional examples of sleep hypnograms generated using SleepVST from video data are shown in Figure 4.13 for a range of Cohen's $\kappa$ agreement values.




(a) Cohen's $\kappa = 0.84$ between model and expert labels.




(b) Cohen's $\kappa = 0.64$ between model and expert labels.




(c) Cohen's $\kappa = 0.39$ between model and expert labels.

Figure 4.13. Example four-class sleep hypnograms from the OSV dataset for various model–expert Cohen's $\kappa$ agreement values. Within each subfigure, the top hypnogram shows expert labels annotated using signals from the vPSG recording, and the bottom hypnogram shows labels automatically generated from near-infrared video using SleepVST.

### 4.A.4 SleepVST Architecture Details

**Convolutional encoder.** Each convolutional layer, denoted 'Kx1 Conv, M' in Figure 4.4, consisted of 1D convolutions with kernel size K, stride 1, and M output channels, followed by batch normalisation [96], and ReLU activation.

**Transformer encoder.** The parameters of our transformer encoder are detailed in Table 4.8. These values were informed by the original design of Vaswani *et al.* [228]. To improve training stability, we employed pre-layer normalisation [250] within each transformer encoder layer.

Table 4.8. SleepVST transformer encoder parameters.

| Architecture Parameter | Value |
|---|---|
| Encoder layers | 6 |
| Self-attention heads | 8 |
| Encoder dropout probability | 0.1 |
| MLP size | 512 |

**Training.** Each training run was performed using a single NVIDIA A10 GPU with 24 GB RAM. Using a sequence length of two hours (N=240) and our default architecture, we used a batch size of 128, the largest power of two that could fit on the GPU. We employed early-stopping to terminate training once there had been no improvement in the validation loss for three consecutive epochs, restoring the model checkpoint that achieved the minimum value. The training run which produced our best model took 6.4 h. All models were implemented using the PyTorch [160] framework.

### 4.A.5 Transformer Output Tiling

Because of the quadratic complexity of the original Transformer [228], we used an input sequence length of N=240 epochs, i.e. two hours, to the SleepVST model. This is much shorter than the sleep recordings within each dataset, which typically last between 8 and 12 hours. To apply SleepVST to these longer sequences, we re-applied the model at up to 30-minute steps, resulting in multiple outputs for

timesteps away from the start and end of the recording. Using this approach, the model can be applied to 10 hours of waveform data in $\approx 0.8$ s.

When directly applying the pre-trained SleepVST model, e.g. to SHHS and MESA test sets, we took the mode of the overlapping classifications as the output classification. This is illustrated in Figure 4.14.

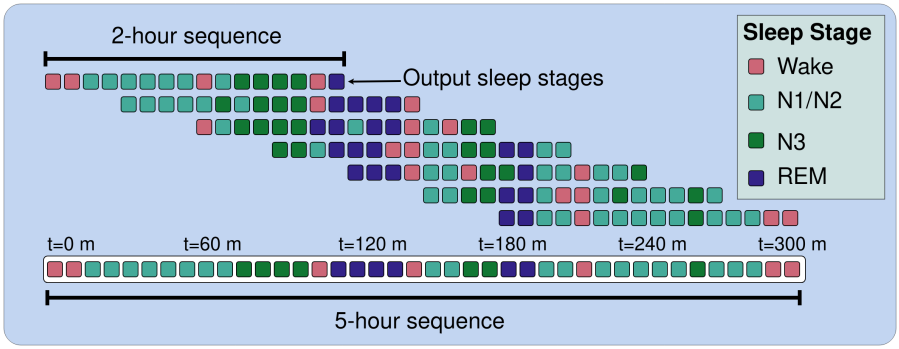



Figure 4.14. **Classifying longer input sequences using SleepVST.** After applying the model to overlapping two-hour sub-sequences, the modal classification at each timestep is used as the output classification.

Similarly, when using SleepVST as a feature extractor on video-derived waveforms, we re-applied the model at intervals to produce overlapping two-hour sequences of feature vectors. For timesteps with multiple feature vectors, we used the feature vector which was closest to the middle of a sequence. This is illustrated in Figure 4.15.

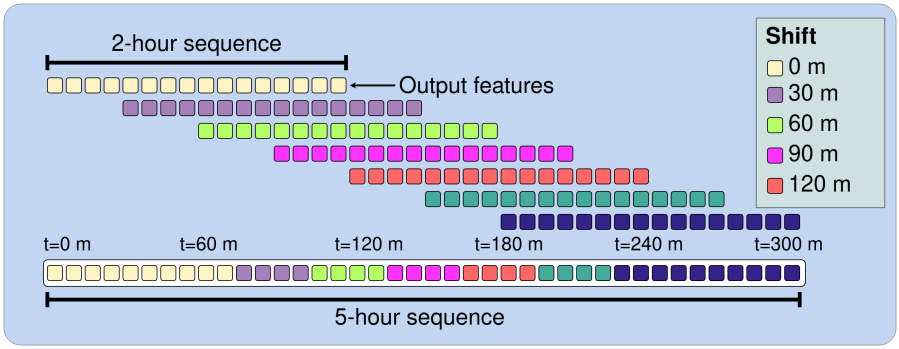



Figure 4.15. **Extracting features from longer input sequences using SleepVST.** The model is applied to two-hour input sub-sequences, producing overlapping output feature sequences. For each timestep, we use the feature vector which is closest to the middle of a sequence.

### 4.A.6 Optical Flow Estimation

We used the Dense Inverse Search algorithm [111] to calculate the optical flow field at a frequency of 4 Hz from the homography-transformed frames, following the same procedure as [37].

### 4.A.7 Cohen's Kappa Calculation

From a confusion matrix $M \in \mathbb{R}^{C\times C}$ with elements $m_{ij}$, the function $K : \mathbb{R}^{C\times C} \to \mathbb{R}$ which calculates Cohen's $\kappa$ statistic is given by:

$$K(M) = 1 - \frac{\sum_{i,j} w_{ij} m_{ij}}{\sum_{i,j} w_{ij} e_{ij}} \tag{4.9}$$

where $w_{ij}$ and $e_{ij}$ are defined as follows:

$$w_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{otherwise} \end{cases} \tag{4.10}$$

$$e_{ij} = \frac{(\sum_{i'} m_{i'j}) \cdot \left(\sum_{j'} m_{ij'}\right)}{\sum_{i'j'} m_{i'j'}} \tag{4.11}$$

### 4.A.8 SHHS Test Set

The 500 randomly sampled participant IDs used to form our test set can be found in the Appendix of our arxiv submission: `https://arxiv.org/abs/2404.03831`

# Part II

# Sleep Staging from Wearables

# 5
# Unified Multi-Modal Sleep Staging

## Preface

This chapter contains the paper "wav2sleep: A Unified Multi-Modal Approach to Sleep Stage Classification from Physiological Signals". This work was presented at the 4$^{\text{th}}$ Machine Learning for Healthcare (ML4H) Symposium 2024 in Vancouver. I was the lead author on the paper, devised the proposed methods, conducted all of the experiments, and wrote the initial manuscript. A full Statement of Authorship can be found in Appendix A.

The preceding chapters have shown that deep learning-based methods can attain superior classification performance in the task of automated sleep staging, but are data-intensive. To adapt these models to scarce video data, Chapters 3 and 4 used the well-established technique of transfer learning: training a model using a data-abundant source domain, then fine-tuning it on a smaller amount of data from the target domain. Returning to contact sensor data, this chapter explores an alternative approach to address this problem, introducing a novel architecture that can be jointly trained on heterogeneous datasets, i.e. where the availability of input modalities may vary. After training, this single, unified model can be applied to any subset of the modalities seen during training. This leads to improved performance across a range of test-time modalities, such as ECG, PPG and respiratory signals.

**Data and Code Availability** All datasets used in this work are available via the National Sleep Research Resource (NSRR, [260]). Training code, model weights, and processing pipelines for all datasets can be found at the following URL: https://github.com/joncarter1/wav2sleep.

## Contents

**5.1 Introduction . . . . . 86**
**5.2 Background and Motivation . . . . . 88**
**5.3 Model Architecture . . . . . 91**
5.3.1 Signal Encoders . . . . . 91
5.3.2 Epoch Mixer . . . . . 92
5.3.3 Sequence Mixer . . . . . 92
5.3.4 Advantages . . . . . 93
**5.4 Experimental Set-up . . . . . 93**
5.4.1 Datasets and Preprocessing . . . . . 93
5.4.2 Model training . . . . . 95
5.4.3 Model hyper-parameters . . . . . 96
5.4.4 Stochastic masking . . . . . 96
**5.5 Results and Discussion . . . . . 97**
5.5.1 Cross-modal learning . . . . . 98
5.5.2 Varying modalities . . . . . 99
5.5.3 Stochastic masking . . . . . 103
5.5.4 Comparison with prior work . . . . . 103
**5.6 Conclusions . . . . . 104**
**5.A Chapter Appendix . . . . . 105**
5.A.1 Additional results . . . . . 105
5.A.2 Dataset processing . . . . . 105
5.A.3 Model design . . . . . 106
5.A.4 Model training . . . . . 108

# 5.1 Introduction

Quantitative analysis of sleep stages is important for applications such as the diagnosis of sleep disorders, the validation of sleep disorder medications, and the discovery of sleep biomarkers. This is typically done through a polysomnography (PSG) study following the American Academy of Sleep Medicine (AASM) guidelines [94], during which a subject will undertake one or more nights of sleep whilst being continuously monitored using a variety of sensors.

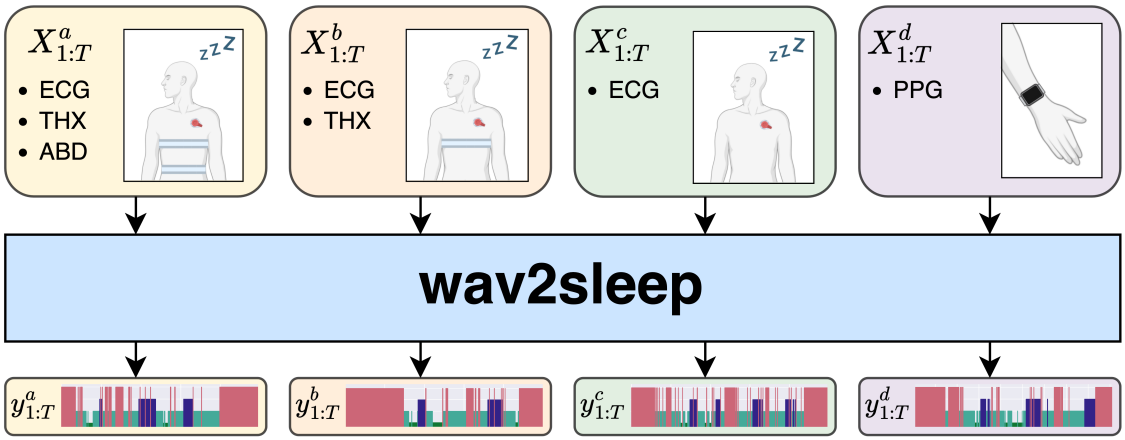



Figure 5.1. **Overview of wav2sleep**. The model operates on *sets* of time-series signals $\boldsymbol{X}_{1:T}$ to classify sleep stage sequences $y_{1:T}$. This enables it to be jointly trained on heterogeneous datasets, with different available input signals, which are especially common in the healthcare domain. At inference time, the same model can be applied to any subset of the signals seen during training.

After the study, a human expert will review the recording and assign each 30-second window of data (a sleep epoch) to one of 5 discrete sleep states defined by the AASM: Wake, N1 (light), N2 (intermediate), N3 (deep), or rapid eye movement (REM) sleep. This is traditionally done using signals such as the electroencephalogram (EEG), electrooculogram (EOG), and chin electromyogram (EMG), which are measured using electrodes attached to the subject's head.

The goal of sleep stage classification (sleep staging) algorithms is to automate this process using one or more input signals. Accurate sleep staging from inputs such as the photoplethysmogram (PPG), electrocardiogram (ECG), abdominal (ABD) and/or thoracic (THX) respiratory signals (see Figure 5.1) is of particular interest since acquiring these signals is typically more comfortable for the subject and easier to set up than a full PSG study.

Sleep staging models have commonly been trained and evaluated using one or more fixed input modalities. However, the recorded signals often have high mutual information, meaning that information available from one input modality may be useful for learning to classify sleep stages from another.

To better leverage this information, we introduce wav2sleep, a unified model for sleep stage classification that operates on sets of physiological signals. During

training, the model can use all available input signals from each night of data, enabling it to be jointly trained across heterogeneous datasets, where the available signals may differ between recordings. This trait is particularly valuable for our problem domain of sleep staging since the set of recorded signals often varies, and many are often discarded due to poor signal quality.

After jointly training on multiple modalities, our model can use any subset of the original input signals during inference. We show that this unified approach outperforms existing methods across several test-time input modalities and datasets.

To summarise, our contributions are as follows:

- We introduce wav2sleep, a novel deep learning architecture for multi-modal time-series that operates on sets of signals. This enables joint training on heterogeneous datasets, where the availability of signals can vary.

- We introduce a simple but effective stochastic masking procedure during training, which enables generalisation to an arbitrary subset of signals at test time.

- After training, we show that our single, unified model outperforms prior methods for sleep staging across a range of modalities. For example, outperforming existing ECG-based methods when using only the ECG at test time.

## 5.2 Background and Motivation

Prior work has shown that deep learning models can classify stages of sleep from brain signals recorded using the EEG with expert-level accuracy [163]. However, EEG signals are typically measured using electrodes attached to the patient's scalp. To overcome this, other work has investigated classifying sleep stages using alternative, less obtrusive modalities such as the ECG [203], PPG [110], and combinations of cardiac and respiratory signals [17].

**Two-step learning.** Automatic sleep staging from modalities such as the ECG or PPG is possible because these signals encode measures of physiological activity, such as heart rate variability (HRV), and these are known to be predictive of the sleep stage [195]. Sleep staging from these signals can therefore be formulated as a two-step learning problem. First, we must find a mapping $f : \boldsymbol{x} \mapsto \boldsymbol{z}$ from the input signals $\boldsymbol{x}$ to relevant physiological features $\boldsymbol{z}$, then a mapping $g : \boldsymbol{z} \mapsto y$ from physiological features $\boldsymbol{z}$ to sleep stages $y$. For models trained end-to-end to classify sleep stages from input signals, these mappings $f$ and $g$ are jointly and implicitly learnt by a deep neural network. This has been shown to outperform methods where the physiological features $\boldsymbol{z}$ are derived using a manual feature engineering approach [110], i.e. where the mapping $f$ is explicit and human-designed.

**Transfer learning.** Sleep staging from modalities such as the PPG is of particular interest since they can be measured from ubiquitous wearables such as smartwatches [42]. However, PPG signals are less common in historical PSG datasets [174], hindering the use of deep learning methods which are notoriously data-hungry. To overcome this issue, transfer learning has been used to improve the performance of PPG-based models, by first pre-training using ECG data [110, 174].

Transfer learning approaches from ECG to PPG signals work because these signals have high mutual information and morphological similarity. For example, both signals encode HRV and respiratory rate variability (RRV) information. This fact is explicitly used by Radha *et al.* [174], who first pre-train a model using HRV features derived from the ECG, then fine-tune using HRV features derived from PPG data. Using the two-step learning formulation introduced at the start of this section, this approach can be thought of as explicitly mapping each input signal $\boldsymbol{x}_i$ to a modality-agnostic feature space $\boldsymbol{z}$ via a modality-dependent mapping $f_i$.

**Modality-agnostic learning.** Ideally, modality-agnostic relationships such as the link between HRV and deep sleep [195], or between RRV and REM sleep [102], should not need to be learnt from a specific modality, a fact which is already partially exploited by transfer learning approaches. However, transfer learning is

prone to the problem of catastrophic forgetting [103] i.e. where after training on new data, information learnt from the old dataset is 'forgotten'. Hence, even if our end goal is to produce a specialised model for a particular modality, we will show that it is beneficial to use a model that is jointly trained across multiple modalities, rather than directly training or using a transfer learning approach.

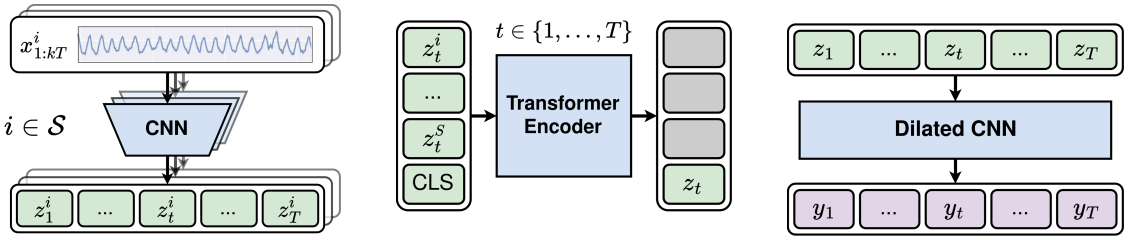
**Multi-modal learning.** Multi-modal sleep staging methods have varied in their approaches to combining cross-modal information. For example, the model proposed by Chambon *et al.* [40] independently turns each input signal into features before passing them to a classifier (late fusion). In contrast, models such as SeqSleepNet [164] concatenate modalities at the input level (early fusion). In the middle ground, separately extracting features from each modality for each 30-second sleep epoch, then fusing and jointly modelling sequential information, is also a common approach, e.g. via concatenation [17, 38] or cross-modal attention [169, 237].

For EEG and EOG signals, Kontras *et al.* [108] handle heterogeneity by simultaneously training both uni-modal and multi-modal models using an alignment loss, with the uni-modal model used when only one signal is available. This is shown to improve uni-modal performance at test time. Alternatively, MaskSleepNet [261], designed for EEG, EOG and/or EMG signals, handles heterogeneity by zero-padding the EOG and/or EMG input signals, which is done randomly during training.

**Objective.** Our aim is to learn a modality-agnostic intermediate representation of the input signals because we hypothesise this can enable more robust representation learning, improving sleep staging performance. However, we want to maintain the ability to perform end-to-end deep learning from the raw input signals, since this has led to improved performance not just in sleep staging, but across numerous problem domains [67]. Finally, we desire a model that can be jointly trained across heterogeneous datasets, improving the diversity of input data, and avoiding the problem of catastrophic forgetting. In Sections 5.3 and 5.4, we describe the wav2sleep model which addresses these challenges, and discuss its advantages over

prior methods. In Section 5.5, we empirically validate its effectiveness in sleep staging across a range of test-time modalities.

## 5.3 Model Architecture




(a) Signal Encoders (b) Epoch Mixer (c) Sequence Mixer

Figure 5.2. **wav2sleep architecture for sets of signals.** $(a)$ Each input signal $x^i_{1:kT}$ from modality $i \in \mathcal{S}$ is passed to a CNN to form a sequence of feature vectors $\boldsymbol{z}^i_{1:T}$. $(b)$ For each time-step $t$, a transformer encoder turns the set of features into a single aggregate feature vector $\boldsymbol{z}_t$ using a CLS token [56]. $(c)$ A dilated CNN mixes sequential information to classify sleep stage output sequences $y_{1:T}$

In this section, we describe the architecture of the wav2sleep model, which turns sets of time-series signals spanning multiple hours into sleep stage classifications for each 30-second sleep epoch. The model architecture, illustrated in Figure 5.2, consists of three high-level components:

1. **Signal Encoders**, which independently extract features for each signal in the input set $\boldsymbol{X}_{1:T}$.

2. **Epoch Mixer**, which fuses cross-modal information into a unified representation $\boldsymbol{z}_t$ for each sleep epoch.

3. **Sequence Mixer**, which mixes information temporally to classify sleep stages $y_{1:T}$.

### 5.3.1 Signal Encoders

The model first turns the set of continuous 1D input signals $\boldsymbol{X}_{1:T} = \{x^i_{1:kT} | \, i \in \mathcal{S}\}$ into a set of feature vector sequences $\boldsymbol{Z}_{1:T} = \{\boldsymbol{z}^i_{1:T} | \, i \in \mathcal{S}\}$, where $\boldsymbol{z}^i_t$ denotes the feature vector for modality $i$ for sleep epoch $t$, $k$ denotes the relative sampling

rate of each signal, and $\mathcal{S}$ denotes the set of available modalities e.g. ECG and PPG signals. We use separate CNN encoders for each input modality, which follow the design of the early layers of SleepPPG-Net [110]. These consist of a stack of residual layers [78], each containing three convolutional layers followed by a max pooling layer to downsample the signal by a factor of 2. The residual layers are followed by a reshape operation and a time-distributed dense layer to produce the sequence of feature vectors $\boldsymbol{z}^i_{1:T}$.

### 5.3.2 Epoch Mixer

Having independently transformed each modality $i$ into a sequence of feature vectors $\boldsymbol{z}^i_{1:T}$, we next fuse information from the set of modalities to provide a single unified representation $\boldsymbol{z}_t$ for each sleep epoch i.e. to complete the mapping $f$ described in Section 5.2. We use a transformer encoder [228] to do this, providing the transformer with an extra learnable vector, i.e. a CLS token [56, 59], and using the output at that position as our unified feature vector. This design straightforwardly handles a varying number of input modalities during training and inference whilst keeping the dimensionality of the fused feature sequence $\boldsymbol{z}_{1:T}$ fixed.

### 5.3.3 Sequence Mixer

The feature vectors $\boldsymbol{z}_{1:T}$ are passed to the sequence mixer, which mixes sequential information to produce sleep stage outputs $y_{1:T}$. This is a desirable property since sleep exhibits long-range time-series structures such as sleep cycles [161]. We use a dilated CNN design as previously used by Sridhar *et al.* [203] and Kotzen *et al.* [110]. This consists of multiple blocks of dilated convolutional layers where the dilation doubles at each layer, meaning that the size of the model's receptive field increases exponentially with network depth.

### 5.3.4 Advantages

Returning to the two-step learning formulation ($f$ and $g$) introduced in Section 5.2, the wav2sleep architecture has two key advantages:

1. Because it operates on *sets* of input signals, the model can be trained on heterogeneous datasets, increasing the variety of data available in terms of both the input modalities (for learning $f$) *and* physiology (for learning $g$).

2. By training on all available modalities *jointly*, this should lead to more robust learning in the presence of noise, by avoiding shortcut learning when one or more modalities are corrupted.

Using a unified model also has practical advantages, since only a single model needs to be trained, validated and deployed, reducing operational complexity for real-world applications.

## 5.4 Experimental Set-up

### 5.4.1 Datasets and Preprocessing

We use 7 PSG datasets available from the National Sleep Research Resource [260]: SHHS [170], MESA [44], CFS [177], MROS [29], CHAT [137], CCSHS [180], and WSC [256]. Demographic information for the datasets used is provided in Table 5.1. Collectively, these datasets contain over 15,000 pairs of overnight polysomnography recordings and expert-annotated sleep stages. Notably, there is significant variation in patient demographics. For example, the SHHS, MESA and WSC datasets are mostly comprised of recordings from older adults with high apnea-hypopnea indices (sleep-disordered breathing). In contrast, the CCSHS and CHAT datasets both contain PSG recordings from children. Joint training across all datasets exposes the model to a wider variety of contact sensors (makes, models etc.) and individual physiological variations.

There is also variation in the availability of signals between the datasets. For example, recordings from MESA and CCSHS, and some from the CFS and CHAT

Table 5.1. Demographics, dataset split sizes, and signal availability for the datasets used.

| Characteristic | SHHS | MESA | WSC | CHAT | CFS | CCSHS | MROS |
|---|---|---|---|---|---|---|---|
| **Demographics**[†] | | | | | | | |
| Age, mean | 65.2 | 69.6 | 59.8 | 7.2 | 41.4 | 17.7 | 78.7 |
| Sex, m:f | 0.88:1 | 0.87:1 | 1.17:1 | 0.94:1 | 0.81:1 | 1.02:1 | 1:0 |
| AHI[‡], mean | 15.2 | 20.4 | 20.0 | 5.5 | 13.2 | 1.5 | 18.3 |
| **Splits**, N | | | | | | | |
| Train | 6441 | 1541 | 1380 | 1132 | 452 | 272 | 0 |
| Validation | 500 | 100 | 250 | 100 | 50 | 50 | 0 |
| Test | 1000 | 200 | 500 | 200 | 100 | 100 | 1000 |
| **Signals** | | | | | | | |
| ECG/ABD/THX | 7941 | 1841 | 2130 | 1432 | 602 | 422 | 1000 |
| PPG | 0 | 1841 | 0 | 1139 | 284 | 422 | 0 |

[†]Calculated from NSRR variables (nsrr_age, nsrr_sex, nsrr_ahi_hp3u). [‡]AHI: Apnea-Hypopnoea Index.

datasets, contain a PPG signal, but recordings from the other datasets do not. Where available, we used the ABD and THX respiratory signals, the PPG signal, and the ECG signal from each recording.

**Dataset splits.** Although numerous prior works have explored the problem of sleep staging on the datasets used, there are no widely-established fixed training, validation and test partitions. We therefore establish new splits for all data sets, excluding nights that have not been annotated with multiple sleep stages as done in prior work [165]. For datasets that contain multiple recordings from a single participant, we ensured that no participant appeared in both the test set and either the training or validation sets. No other exclusion criteria–such as signal quality heuristics [97]–were explicitly used since one of the key aims of training on multiple modalities jointly is to improve robustness to noise on any particular channel.

The size of our training, validation and test set splits for each dataset are listed in Table 5.1. These were chosen to be in line with those used in prior work, e.g. [203]. Our splits were carefully constructed to additionally allow evaluation on the aggregated test set proposed by Jones *et al.* [97], which uses multiple PSG datasets to create a test set that approximately matches the 2022 US census demographics. Throughout the remainder of this paper, we refer to this as the

'Census' test set. More detail on the construction of our training, validation and test sets is provided in Section 5.A.2.

**Preprocessing.** We minimally processed all signals using a similar process to that described by Kotzen *et al.* [110], padding or truncating each recording to 10 h (i.e. sequence length $T = 1200$), re-sampling each signal to the same frequency across recordings, and applying unit normalisation. The ECG and PPG signals were resampled such that each 30-second sleep epoch consisted of $k = 1024$ data points ($\approx 34$ Hz), which simplifies temporal alignment during pooling operations within the convolutional layers of the signal encoders. Since respiratory signals are generally sampled at a lower frequency during PSG recordings (e.g. 5-10 Hz in SHHS), the ABD and THX signals were resampled to a lower frequency of $k = 256$ data points per sleep epoch ($\approx 8$ Hz), reducing the computational and memory requirements of the model during training and inference.

### 5.4.2 Model training

All models were trained to minimise the cross-entropy loss between expert-annotated sleep stages and model outputs using the AdamW optimiser [134] with a batch size of 16 and weight decay of $10^{-2}$. For the learning rate schedule, we used a linear warm-up of 2000 steps to a maximum learning rate $\epsilon = 10^{-3}$ followed by an exponential decay to zero. Training continued until there was no decrease in the loss on the validation set for 5 epochs, which typically required around 30 epochs in total. Further training details can be found in Section 5.A.4. The checkpoint which resulted in the lowest validation loss was restored for evaluation. Model hyper-parameters were tuned using the validation sets before evaluation on the test sets took place.

**Augmentation.** As noted by Jones *et al.* [97], signals such as the ECG are sometimes inverted due to electrodes being connected the wrong way around. To improve robustness, all signals were randomly inverted (multiplied by -1) with a 50% probability during training.

### 5.4.3 Model hyper-parameters

Hyper-parameters for the wav2sleep model are listed in Table 5.2. In each signal encoder, the number of residual layers, and the number of channels in each layer, were chosen so that the resulting feature dimension is independent of the relative sampling rate $k$. For simplicity, we retained this feature dimension ($\dim(\boldsymbol{z}_t^i) = \dim(\boldsymbol{z}_t) = 128$) throughout the remainder of the model. Additional architecture details can be found in Section 5.A.3.

Table 5.2. wav2sleep model hyper-parameters.

| Hyper-parameter | Value |
|---|---|
| **Global** | |
| Feature dimension | 128 |
| Activation function | GELU[†] |
| Dropout | 0.1 |
| **Signal Encoders** | |
| Kernel size | 3 |
| Channels ($k = 256$) | (16,32,64,64,128,128) |
| Channels ($k = 1024$) | (16,16,32,32,64,64,128,128) |
| **Epoch Mixer** | |
| Transformer layers | 2 |
| Hidden dimension | 512 |
| Attention heads | 8 |
| **Sequence Mixer** | |
| Dilated blocks | 2 |
| Kernel size | 7 |
| Dilations (per block) | (1,2,4,8,16,32) |

[†]Hendrycks and Gimpel [82].

### 5.4.4 Stochastic masking

During training, to handle differences in the available modalities within a batch, we padded unavailable signals and added a mask to the attention matrices of the epoch mixer. To aid test-time generalisation to a subset of modalities, we randomly sampled a subset of the available signals for each recording via additional masking of the attention matrix. Where available, the input signals were masked

with the following probabilities:

$$p(m_{\text{ABD}}) = 0.7 \quad p(m_{\text{THX}}) = 0.7$$
$$p(m_{\text{ECG}}) = 0.5 \quad p(m_{\text{PPG}}) = 0.1$$

These values were intuitively chosen so that the higher frequency ECG and PPG signals were less likely to be masked, and to increase the prevalence of the scarcer PPG signal.

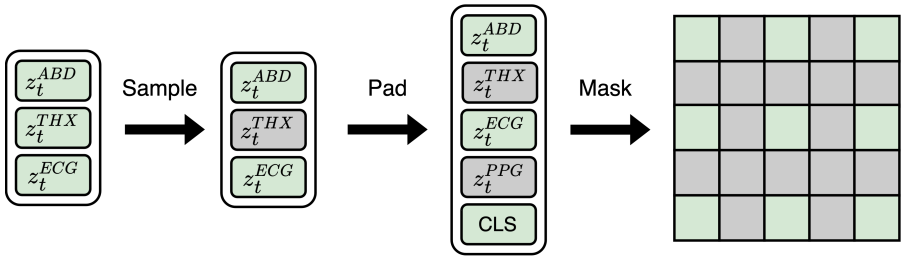



Figure 5.3. **Stochastic masking.** During training, we sample a random subset of the available modalities for each night of data. To retain a fixed batch shape, we pad unavailable modalities and apply a mask to the self-attention matrices of the epoch mixer.

Our stochastic masking process draws inspiration from masked modeling approaches e.g. [15, 79], and is similar to hierarchical channel sampling (HCS, [21]) from the visual domain. However, the practical implementation of HCS requires all samples within a batch to have the same available channels, to retain a fixed batch shape during training. Our approach (illustrated in Figure 5.3) allows heterogeneity *within* batches, which simplifies training on heterogeneous datasets. During inference, we simply pass only the available signals to the transformer i.e. there is no need for any padding or manipulation of the attention mechanism to handle different numbers of input signals at test time.

## 5.5 Results and Discussion

In this section, we report the performance of our model in four-class sleep staging, merging N1 and N2 into a single 'Light' sleep class as commonly done in prior work. We report total Cohen's $\kappa$ ($\kappa_T$) and accuracy ($\text{Ac}_T$) calculated over all sleep epochs in each test set. All results are averages over three training runs using

different random seeds. Our full set of results for all dataset-modality combinations evaluated can be found in Table 5.6 of Section 5.A.1.

### 5.5.1 Cross-modal learning

In Table 5.3, we compare the performance of three approaches to PPG-based sleep staging:

1. Direct training on (scarce) PPG signals.
2. Transfer learning from ECG to PPG signals.
3. Joint training on all available modalities.

We additionally compare with a re-implementation of SleepPPG-Net [110], trained using the same splits and learning procedure as our model. Using transfer learning ($\mathcal{S}_{\text{Train}} = \text{ECG} \rightarrow \text{PPG}$), we pre-train using the ECG signal, then fine-tune using the PPG signal, resuming the learning rate schedule. Across datasets, we find that our joint training approach with stochastic masking consistently leads to better performance than either direct training or transfer learning for the scarce PPG modality.

Table 5.3. Performance ($\kappa_T$) for $\mathcal{S}_{\text{Test}} = \text{PPG}$.

| | | Dataset | | | |
|---|---|---|---|---|---|
| Model | $\mathcal{S}_{\text{Train}}$ | MESA | CHAT | CFS | CCSHS |
| SleepPPG-Net | PPG | 0.713 | 0.757 | 0.731 | 0.803 |
| | ECG→PPG | 0.724 | 0.767 | 0.746 | 0.808 |
| wav2sleep | PPG | 0.728 | 0.777 | 0.751 | 0.817 |
| | ECG→PPG | 0.732 | 0.779 | 0.754 | 0.811 |
| | All[†] | **0.742** | **0.793** | **0.763** | **0.832** |

[†]ABD+THX+ECG (+PPG for MESA, CHAT, CFS, CCSHS)

Similarly, Table 5.4 compares the performance of direct and joint training for sleep staging using the (abundant) ECG signal. For the SHHS and WSC datasets, and for the completely held-out MROS dataset, joint training resulted in the best performance. However, for some datasets, we found that joint training without the PPG signal ($\mathcal{S}_{\text{Train}} = \text{No PPG}$) resulted in better ECG-only performance. This

indicates that cross-modal learning from respiratory signals to ECG was able to occur, but that there is a trade-off between learning from ECG and PPG signals for some datasets.

Table 5.4. Performance ($\kappa_T$) for $\mathcal{S}_{\text{Test}} = \text{ECG}$.

| Model | $\mathcal{S}_{\text{Train}}$ | Dataset SHHS | WSC | CFS | Census | MROS |
|---|---|---|---|---|---|---|
| SleepPPG-Net | ECG | 0.722 | 0.671 | 0.762 | 0.765 | 0.712 |
| wav2sleep | ECG | 0.733 | 0.683 | 0.785 | 0.786 | 0.746 |
| | No PPG | 0.738 | 0.686 | **0.789** | **0.792** | 0.748 |
| | All | **0.739** | **0.689** | 0.784 | 0.783 | **0.750** |

†ABD+THX+ECG (+PPG for MESA, CHAT, CFS, CCSHS)

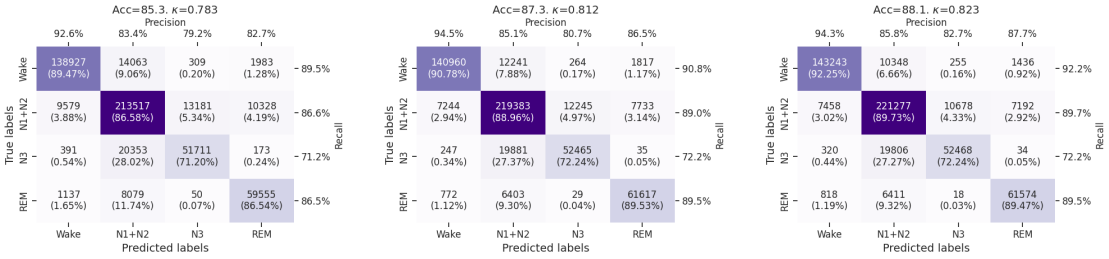
During training, there are a finite number of optimisation steps before the model begins to overfit. This results in a small performance trade-off between different test-time modality combinations depending on the masking parameters, which determine the relative frequency of modalities observed during training. For example, by increasing the PPG masking probability $p(m_{\text{PPG}})$ to 0.2 we found that $\kappa_T$ values slightly decreased by 0.01-0.02 across datasets when using only the PPG at test-time, but increased by around the same amount using just the ECG.

A similar effect was noted in the similar approach of Hierarchical Channel Sampling [21], where performance was best for combinations of channels that were most frequently sampled during training. Our stochastic masking procedure means that ECG-only examples are infrequently sampled during training, accounting for less than 1 example per batch on average for datasets that have all four signals available. Improvements to the stochastic masking procedure, stronger regularisation, and/or a larger batch size may help to address this in future work.

### 5.5.2 Varying modalities

Figure 5.4 shows confusion matrices between expert-annotated sleep stages and model outputs for different test-time modalities using the Census test set, aggregated over all sleep epochs. Here we see that the addition of breathing signals (ABD, THX) is particularly helpful in distinguishing both Wake and REM from Light

(N1+N2) sleep. Using just the ECG and THX signals, we obtain a Cohen's $\kappa$ of 0.812. Whilst caution should be taken when interpreting $\kappa$ values, notably, using the rule-of-thumb proposed by Landis and Koch [118] this corresponds to 'almost perfect' agreement with the expert-annotated sleep stages.




(a) $\mathcal{S}_{\text{Test}} = \text{ECG}$ (b) $\mathcal{S}_{\text{Test}} = \{\text{ECG, THX}\}$ (c) $\mathcal{S}_{\text{Test}} = \text{All}$

Figure 5.4. Confusion matrices for varying $\mathcal{S}_{\text{Test}}$ on the Census test set.

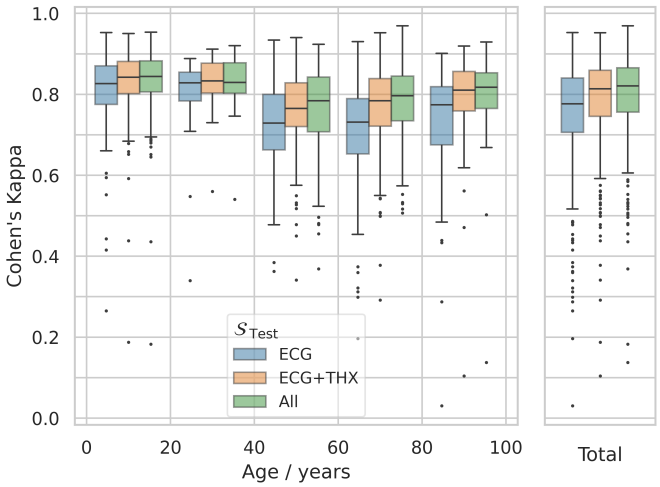



Figure 5.5. Performance against age for varying $\mathcal{S}_{\text{Test}}$ on the Census test set.

In Figure 5.5, we plot the performance of the wav2sleep model for different age ranges and test-time modalities $\mathcal{S}_{\text{Test}}$. We observe good performance across age ranges, and that using more modalities consistently leads to improved performance, particularly by reducing the quantity and severity of outliers. As shown in Figure 5.6, these are often caused by low signal quality, with additional signals providing improved redundancy. However, they can also be caused by specific physiological conditions.

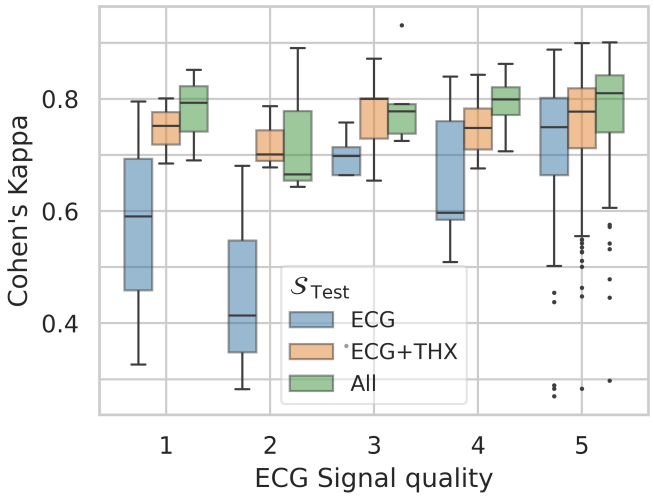



Figure 5.6. Performance grouped by ECG signal quality.

Figure 5.7 shows the performance of wav2sleep on the WSC dataset, stratified by apnea-hypopnea index and cardiac arrhythmia diagnosis. Notably, we found that when using the ECG as the sole input, performance **improved** with apnea severity for subjects with cardiac arrhythmia. This is in contrast to the general trend seen in prior work that the performance of sleep staging models tends to decrease with apnea severity [109]. We hypothesise that, for subjects with arrhythmia, the model may mistake heart rate variability (HRV) caused by arrhythmia for HRV caused by the more common condition of sleep apnea [162]. In turn, this may confound the learnt mapping between physiological features and sleep stages i.e. the mapping $g$ described in Section 5.2.

Figure 5.8 shows example sleep hypnograms generated by the wav2sleep model using different test-time modalities. This night of data[1] corresponds to an elderly male with diagnosed cardiac arrhythmia and mild sleep apnea. Visually, the ECG is of good quality, however, using the ECG as the sole input to the model results in poor agreement with expert-annotated sleep stages. Adding a respiratory signal (e.g. using ECG and THX) appears to remedy this issue, enabling the model to correctly distinguish apnea-induced variability.

[1]Session ID: wsc-visit2-12529-nsrr

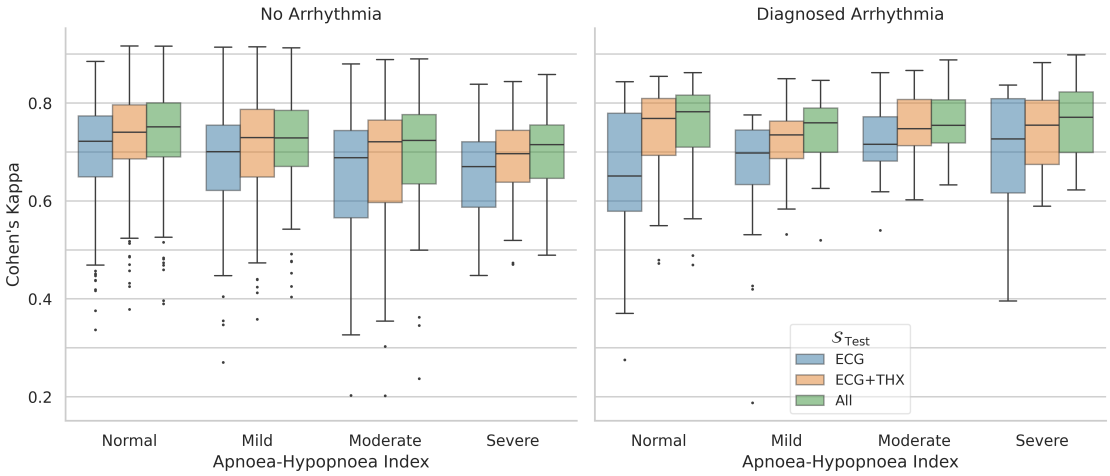



Figure 5.7. **Performance of wav2sleep on the WSC test set, grouped by apnea severity and arrhythmia**. For subjects with arrhythmia, and using only the ECG at test-time, performance *improves* with apnea severity. This is in contrast to the trend that performance decreases with apnea severity seen using other modalities and for subjects without diagnosed arrhythmia.

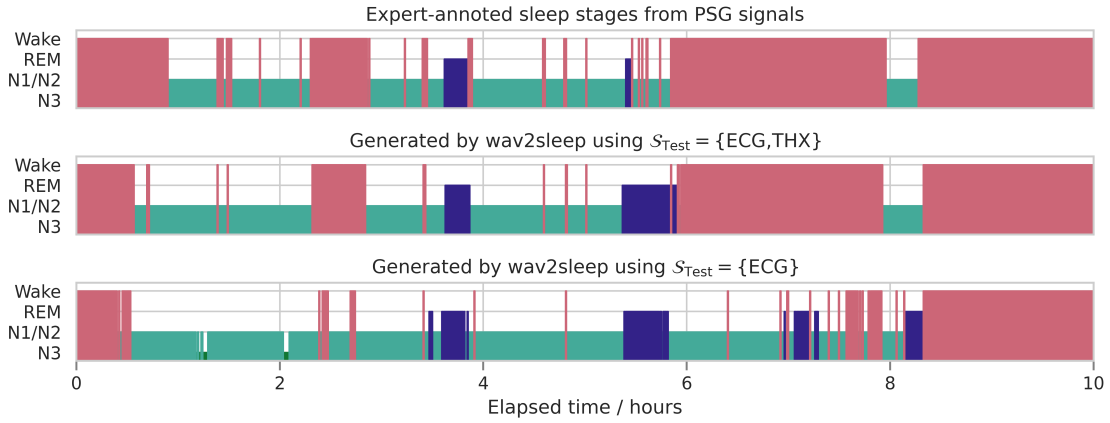



Figure 5.8. **Example sleep hypnograms for a subject with diagnosed cardiac arrhythmia.** (top) Annotated by a human expert using the PSG recording. (middle) Produced by the wav2sleep model using ECG and THX signals ($\kappa_T = 0.74$). (bottom) Produced by the wav2sleep model using the ECG signal ($\kappa_T = 0.19$).

This highlights how, during real-world deployment, the set of input modalities (contact sensors) may need to be chosen in a patient-specific manner to ensure an expected level of accuracy in the presence of physiological confounders. Using a single, unified model such as wav2sleep can help simplify such a process.

### 5.5.3 Stochastic masking

Figure 5.9 shows the performance of wav2sleep for various dataset-modality combinations with and without stochastic masking during training. Here we can see that stochastic masking is essential for generalisation to subsets of modalities at test-time, whilst maintaining equivalent performance when using all modalities.

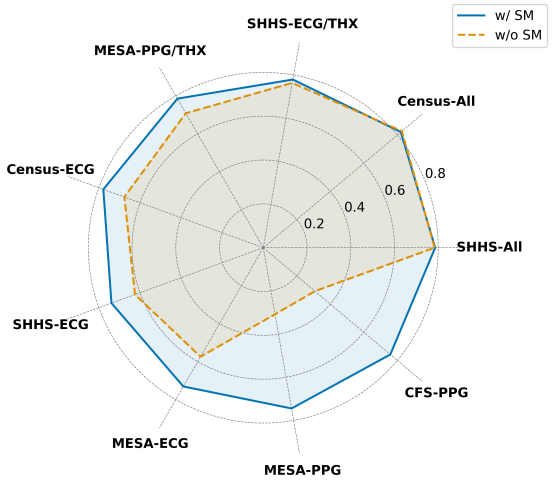



Figure 5.9. Performance ($\kappa_T$) of wav2sleep for various dataset-modality combinations with and without stochastic masking (SM) during training.

### 5.5.4 Comparison with prior work

In Table 5.5, we compare the performance of wav2sleep after training on all modalities with prior models trained on specific modalities. We follow the exclusion criteria of Jones *et al.* [97] and compare against prior work that has explicitly reported the use of distinct training, validation and test sets. Across multiple datasets and combinations of test-time modalities, the wav2sleep model outperforms existing methods for sleep staging from cardio-respiratory signals.

Table 5.5. Comparison of cardio-respiratory sleep staging methods for different test-time modalities $\mathcal{S}_{\text{Test}}$.

| Dataset | $\mathcal{S}_{\text{Test}}$ | Method | $\kappa_T$ | $\text{Ac}_T$ |
|---|---|---|---|---|
| SHHS | ECG, THX | Bakker *et al.* [17]† | 0.64 | 76.7 |
| | | Carter *et al.* [38] | 0.75 | 83.0 |
| | | wav2sleep | **0.78** | **85.0** |
| | ECG | Sridhar *et al.* [203] | 0.66 | 77.0 |
| | | wav2sleep | **0.74** | **82.3** |
| MESA | PPG, THX | Bakker *et al.* [17]† | 0.68 | 79.8 |
| | | wav2sleep | **0.78** | **86.2** |
| | ECG, THX | Carter *et al.* [38] | 0.77 | 85.2 |
| | | wav2sleep | **0.78** | **86.1** |
| | ECG | Sridhar *et al.* [203] | 0.69 | 80.0 |
| | | wav2sleep | **0.73** | **82.8** |
| Census | ECG | Jones *et al.* [97]‡ | 0.77 | - |
| | | wav2sleep | **0.78** | **84.8** |

Additional model inputs: †Nasal airflow,‡age and sex.

## 5.6 Conclusions

In this paper, we have introduced wav2sleep, a deep learning model for automated sleep stage classification that can operate on a variable number of input modalities during training and inference. After joint training on over 10,000 nights of publicly available data from six heterogeneous datasets, this single, unified model leads to improved performance compared to direct training and transfer learning methods across a range of test-time modalities and datasets. Our work further improves the accuracy of sleep staging across a range of important modalities, such as ECG, PPG and respiratory signals, bringing accurate, low-cost sleep monitoring from less obtrusive contact sensors closer to clinical practice.

**Future Work.** We have focused on learning from cardio-respiratory signals since sleep staging from these modalities is of particular interest. However, using additional signals such as the EEG may help to further improve the quality of the learnt representations. Finally, the generalised architecture of wav2sleep, particularly the ability to jointly train it on heterogeneous, multi-modal time-series, means that it could be used to complement unsupervised approaches, e.g. [215].

# 5.A Chapter Appendix

## 5.A.1 Additional results

After joint training on all datasets and input modalities ($\mathcal{S}_{\text{Train}}$ = All), the performance of the wav2sleep model for different test-time modalities $\mathcal{S}_{\text{Test}}$ is listed in Table 5.6.

Table 5.6. wav2sleep performance ($\kappa_T$) for different test-time modalities.

| | Dataset | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\mathcal{S}_{\text{Test}}$ | SHHS | MESA | WSC | CHAT | CFS | CCSHS | MROS | Census |
| All[†] | 0.786 | 0.796 | 0.737 | 0.836 | 0.803 | 0.857 | 0.805 | 0.821 |
| ECG,THX | 0.779 | 0.783 | 0.728 | 0.827 | 0.802 | 0.854 | 0.796 | 0.812 |
| ECG | 0.739 | 0.731 | 0.689 | 0.800 | 0.784 | 0.833 | 0.750 | 0.783 |
| PPG | - | 0.742 | - | 0.793 | 0.763 | 0.832 | - | - |

[†]ABD+THX+ECG (+PPG for MESA, CHAT, CFS and CCSHS)

## 5.A.2 Dataset processing

**Scoring exclusions.** Because of signal quality issues, some of the recordings in each dataset only have binary sleep–wake annotations, rather than full AASM (Wake, N1, N2, N3, REM) sleep stages. For these recordings, *all* sleep stages are typically assigned to the same integer as 'N2' sleep. This means that these labels should not be used for training or evaluation of multi-class sleep staging models. Where available, we used the harmonised 'nsrr_flag_spsw' metadata variable produced by the National Sleep Research Resource to exclude these recordings. Otherwise, we checked for the existence of either N1, N3 or REM sleep labels.

**Construction of test sets.** From the CFS and CHAT datasets, we created our validation and test sets using recordings where the PPG signal was available, to enable evaluation across all combinations of modalities. The remaining recordings (with and without PPG available) were used for training. We used recordings from the non-randomised (single night per participant) arm of the CHAT dataset for our test set. Similarly to [38], we selected 1000 nights for our SHHS test set by randomly choosing 500 participants who participated in both visits. For the WSC

dataset, we selected 2 recordings from 250 participants who had undertaken at least 2 visits to form our test set of 500 recordings; additional recordings from participants in the test set were excluded. Our decision to use test sets with two recordings per person for the SHHS and WSC datasets was taken to maximise data usage while avoiding the same participant appearing simultaneously in both the training and test sets. In future work, these sets could be used for additional analysis such as the variation in performance with age after controlling for identity. To enable evaluation on the census-balanced test set proposed by [97]–which uses recordings from CCSHS, CFS, CHAT, MESA and WSC–we excluded their test set recordings from our training and validation sets.

### 5.A.3 Model design

Here we describe additional experiments and observations that informed the design and hyper-parameters of the wav2sleep model. Hyper-parameter search was informed by the minimum validation loss $\mathcal{L}_{\text{Min}}$ during initial experiments.

**Signal Encoders.** We found that using instance normalisation [218] within the signal encoders and layer normalisation [14] in the sequence mixer improved training stability and performance. Because of our stochastic masking procedure, the number of examples of a signal within a batch will often be much smaller than the actual batch size, increasing the variance of statistics used by the more common approach of batch normalisation [96].

**Epoch Mixer.** We evaluated two designs for the epoch mixer:

1. A small transformer encoder (TE) i.e. our best-performing approach.
2. A linear concatenation and projection layer, handling variation in the available inputs with zero-padding.

The attention-based epoch mixer achieved a lower validation loss and higher Cohen's $\kappa$ values across multiple datasets and modalities as shown in Table 5.7.

Table 5.7. Validation metrics for different epoch mixer designs.

| | | Dataset-modality $\kappa_T$ | | |
|---|---|---|---|---|
| Design | $\mathcal{L}_{\text{Min}}$ | C-ECG | M-PPG | S-ECG |
| Linear | 0.351 | 0.770 | 0.741 | 0.719 |
| TE | 0.349 | 0.773 | 0.743 | 0.723 |

C - Census. M - MESA. S - SHHS

**Sequence Mixer.** We evaluated two designs for the sequence mixer:

1. A dilated convolutional (DCNN) design [110, 203].

2. A transformer encoder (TE) with sliding window attention [24] and rotary positional embeddings [208].

We found that the DCNN design consistently achieved better performance across different modalities (see Table 5.8), and converged after fewer training epochs. The hyper-parameters of the dilated convolutional design have been carefully tuned through extensive hyper-parameter search in prior work [110, 203]. Though we did perform a basic search over transformer hyper-parameters, such as the number of encoder layers and the context length, performing extensive tuning was deemed unnecessary given the results achieved using a convolutional design, and outside the scope of this paper. Using a well-tuned mixture of local and global attention [24] may yet lead to superior performance using a transformer-based architecture, but is left for future work.

Table 5.8. Validation metrics for different sequence mixer designs.

| | | Dataset-modality | | |
|---|---|---|---|---|
| Design | $\mathcal{L}_{\text{Min}}$ | C-ECG | M-PPG | S-ECG |
| TE | 0.355 | 0.764 | 0.724 | 0.706 |
| DCNN | 0.349 | 0.773 | 0.743 | 0.723 |

C - Census. M - MESA. S - SHHS

Finally, it is worth noting that our implementations of both stochastic masking and local attention relied on naive masking of the attention matrix using PyTorch [160]. Using optimised sparse kernels (e.g. from xformers [123]) could

provide significant speed-ups and efficiency gains on modern GPU architectures, making a transformer-based architecture a more attractive option for training on an even larger quantity of data.

### 5.A.4 Model training

Model parameters $\theta$ were found by minimising the unweighted cross-entropy loss between one-hot encoded labels $\boldsymbol{y}_{1:T} \in \mathbb{R}^{C\times T}$ and output probabilities $\boldsymbol{p}_{1:T} \in \mathbb{R}^{C\times T}$. For each night of data, the total cross-entropy loss is given by:

$$\mathcal{L}_\theta(\boldsymbol{y}_{1:T}, \boldsymbol{p}_{1:T}) = -\sum_{i=1}^{C}\sum_{j=1}^{T}(\boldsymbol{y}_{1:T} \odot \log(\boldsymbol{p}_{1:T}))_{ij} \tag{5.1}$$

where $\odot$ denotes the Hadamard product.

**GPU training.** Experiments were performed using a computing cluster containing multiple GPU architectures. Gradient accumulation was used to ensure a consistent effective batch size of 16, using the largest batch size that could fit on the particular GPU(s) used in a given experiment. Using a single NVIDIA A100, the actual batch size was 4 samples, and each epoch took 21 minutes, resulting in an average training time of around 10 hours.

# 6

# Automated Sleep Staging under Uncertainty

## Preface

The preceding chapters have advanced the state-of-the-art in automated sleep staging, both from near-infrared video and from a range of contact sensors. However, the model–expert agreement statistics obtained are not (yet) on a par with expert–expert agreement. This chapter argues that even if such models do not reach expert-level performance, they may still have important clinical applications, so long as they can produce calibrated measures of uncertainty. This chapter addresses a key flaw in the modelling assumptions of existing sleep staging models and introduces a novel model that can provide measures of sleep staging uncertainty over particular points in time, over entire nights, and over sleep metric estimates. This chapter is presented in a traditional thesis format. However, a precursor to this work was presented as an extended abstract at the ICLR 2024 Learning from Time Series for Health Workshop.

## Contents

**6.1 Introduction . . . . . 110**
**6.2 Background and Motivation . . . . . 111**
6.2.1 Why do we need measures of uncertainty? . . . . . 111
6.2.2 How could we use measures of uncertainty? . . . . . 114
6.2.3 What's wrong with existing methods? . . . . . 116
**6.3 Methods . . . . . 118**
6.3.1 SleepDecoder . . . . . 118
6.3.2 Model calibration and posterior sampling . . . . . 120
6.3.3 Measures of uncertainty . . . . . 122
6.3.4 Training and evaluation procedure . . . . . 124
**6.4 Experiments . . . . . 125**
6.4.1 Overall performance . . . . . 125
6.4.2 Uncertainty over sleep stages . . . . . 126
6.4.3 Uncertainty over sleep metrics . . . . . 129
6.4.4 Out-of-distribution detection . . . . . 134
6.4.5 Comparing model-expert and inter-expert agreement . . 136
**6.5 Conclusions . . . . . 137**

# 6.1 Introduction

Over the past three chapters, this thesis has advanced the accuracy of sleep staging algorithms from less obtrusive modalities including wearables and vision. However, the performance of these algorithms still lags behind that of conventional expert-annotated polysomnography, which predominantly uses the EEG and/or EOG as the main input sources. At the time of writing, no medical devices have been cleared by a regulatory body such as the US Food and Drug Administration (FDA) for automated sleep staging from non-EEG sensors.

This chapter argues that a greater focus on *calibrated* uncertainty quantification, i.e. where the output probabilities reflect the empirical probability of correctness, can allow us to harness these algorithms, without necessarily requiring expert-level performance as measured by conventional metrics such as accuracy or Cohen's $\kappa$ statistic. Focusing on the problem of sleep staging from single-channel ECG recordings, this chapter highlights several challenges that wearable sleep staging faces

in real-world deployment and describes how calibrated uncertainty quantification can help to overcome these.

A novel architecture for sleep staging is introduced, which is designed to address a key flaw in the modelling assumptions of most sleep staging methods described in the literature. Techniques from Bayesian deep learning, temperature scaling [72] and deep ensembles [117], are also used to improve model calibration. The resulting model is used to derive simple, principled measures of uncertainty, which are shown to be calibrated, even on completely held-out test sets and for out-of-distribution inputs. Using the multi-scored Dreem Open datasets [71] this model is shown to achieve comparable performance to human experts. Importantly, the model 'knows' when it is not accurate, from its measures of uncertainty which are derived solely from the input data, a property which is essential for the safe deployment of medical devices.

## 6.2 Background and Motivation

This section outlines why a greater focus on uncertainty quantification, rather than just scalar performance metrics such as accuracy or Cohen's $\kappa$, can help to enable clinical applications of sleep staging using wearables. To aid this motivation, we first introduce a statistical observation model for the problem of sleep staging, illustrated in Figure 6.1, to which we will regularly refer back throughout this section. Under this model, we assume that the prevalent AASM scoring rules define a complete mapping $\pi^*$ from polysomnography signals $\boldsymbol{X}_{1:T}^{\text{PSG}}$ to sleep stages $y_{1:T}$.[1] We then use the notation $\pi^e$ to denote the mapping from signals to sleep stages of an expert sleep scorer and $\pi_\theta$ to denote the learnt mapping of an automated sleep staging model, which may use some alternative set of input signals $\boldsymbol{X}_{1:T}^{\text{M}}$ derived from wearables.

### 6.2.1 Why do we need measures of uncertainty?

There are several sources of uncertainty in sleep staging which we briefly review here. For a longer discussion on this topic, we additionally refer the reader to

[1] In practice, these human-designed rules cannot account for all possible variability.

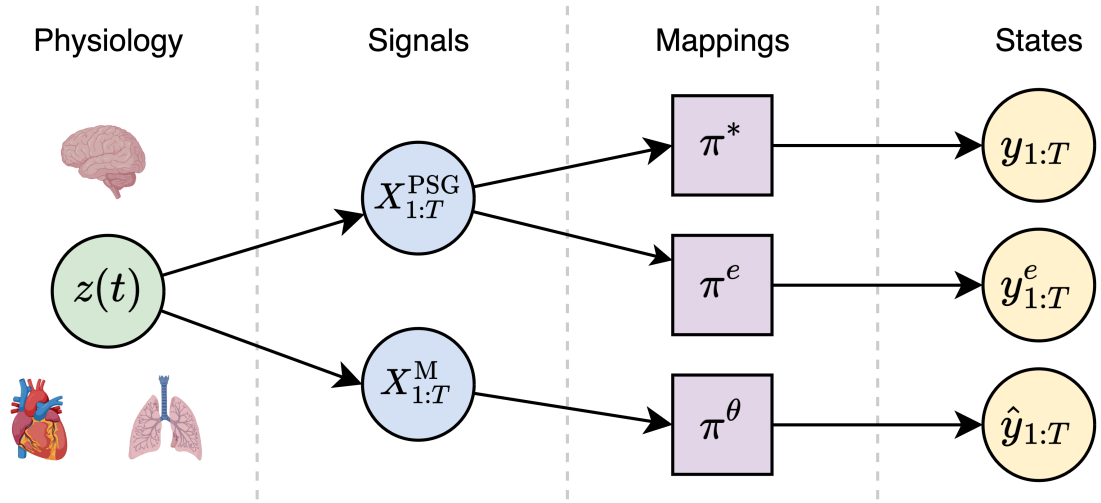



Figure 6.1. **A statistical observation model for automated sleep staging.** The continuous, latent state of the body $z(t)$ drives changes in contact sensor measurements, such as those from polysomnography $\boldsymbol{X}^{\text{PSG}}_{1:T}$. The AASM scoring rules define a mapping $\pi^*$ from PSG signals to sleep stages $y_{1:T}$, which human experts $\pi^e$ attempt to follow. Automated sleep staging methods aim to find a mapping $\pi_\theta$ from some input signal set $\boldsymbol{X}^M_{1:T}$ (often a subset of $\boldsymbol{X}^{\text{PSG}}_{1:T}$) to the same sleep stages.

the work of van Gorp *et al.* [224]. When describing uncertainty and its sources, our choice of language is driven by the recently proposed framework of Bickford Smith *et al.* [201], which divides the uncertainty of a prediction problem into reducible and irreducible components and explicitly acknowledges that measures of uncertainty produced by models such as neural networks are only *estimates* of the true uncertainty. We intentionally avoid the 'aleatoric' vs. 'epistemic' framing of uncertainty prevalent in prior work, which has led to ambiguity and confusion in the machine learning literature [201].

**Sensor limitations.** Like any other contact sensors, wearable modalities such as the ECG and PPG suffer from noise [155]. However, given that the devices that record them usually only measure one or two modalities[2], i.e. $\dim(\boldsymbol{X}^{\text{M}}_{1:T}) < \dim(\boldsymbol{X}^{\text{PSG}}_{1:T})$, they have less redundancy. On the other hand, polysomnography measures brain activity from multiple sites using the EEG, plus other complementary physiological measurements, such as eye movements (EOG) and muscle activity under the chin (EMG), which can all aid the classification of sleep stages. Therefore,

[2]e.g. smartwatches often measure a PPG signal and an accelerometry signal.

when using wearable signals, we should expect a greater proportion of the data to be unscorable due to issues such as signal noise, resulting in irreducible uncertainty.

**Physiology.** Previous chapters have demonstrated that cardio-respiratory activity encodes sleep stage information, enabling the development of models that can classify sleep stages from modalities such as the ECG with high accuracy. However, it may not always be possible to distinguish sleep stages solely from signals such as the ECG, even in the absence of noise. The discrete AASM sleep states are ultimately defined by rules based on characteristic patterns of polysomnography signals $\boldsymbol{X}_{1:T}^{\text{PSG}}$. For example, a 30-second epoch is labelled as N3 (deep) sleep when at least 20% of the epoch consists of slow-wave (0.5–2 Hz) EEG activity above 75 $\mu$V [94]. Referring back to Figure 6.1, automated sleep staging from PSG signals requires a model to approximate the mapping $\pi^*$ that is implicitly defined by the AASM scoring rules. In contrast, sleep staging from alternative signals such as the ECG requires a model to learn about the underlying physiological dynamics $z(t)$ that give rise to the sleep stages. Perfect agreement with expert-annotated stages from PSG would imply that the (somewhat arbitrary) N3 EEG amplitude criteria, a specific measurement of central nervous system (CNS) activity, can always be predicted from the ECG, whose rhythm (the heart rate) is regulated by the autonomic nervous system (ANS). In reality, this is highly implausible, and there will inevitably be irreducible uncertainty caused by this approximation using alternative signals.

Numerous factors can affect CNS-ANS coupling [54], which underpins cardio-respiratory sleep staging. Physical and mental health e.g. [114, 227], medications [157], and factors such as caffeine [22] and alcohol intake [214] may all affect the observed relationship between cardio-respiratory signals and sleep stages. We cannot expect to observe *all* combinations of factors that may affect this relationship during model training before a model is deployed into the real world. Therefore, a model must be able to identify ambiguous or out-of-distribution inputs, to avoid producing incorrect outputs.

**Subjectivity.** Finally, expert-annotated sleep stages are **not** a ground truth, with high inter-expert disagreement [52, 224]. Most polysomnography datasets contain recordings that are scored by a single human expert, and the labels provided by the scorer $y^e_{1:T}$ are assumed to be a perfect 'ground truth' $y_{1:T}$. To address this, Fiorillo *et al.* [62] trained against the 'soft consensus' of multiple expert annotations, which resulted in improved performance over training on single-scored data. However, since manual scoring is a time-consuming process, multi-scored datasets are scarce and so deep learning methods that use much greater quantities of single-scored data still perform much better, e.g. [165, 210].

Figure 6.2 shows three sleep stage hypnograms independently annotated by different experts for the same night of data taken from the Dreem multi-scored datasets [71]. Here we see that there is long-range disagreement between experts regarding the presence of a REM cycle around 23.50 and deep sleep between 02.00 and 03.00. Later in this chapter, we will show that a key modeling assumption of most existing sleep staging models means that they cannot fully characterise this type of long-range uncertainty.

### 6.2.2 How could we use measures of uncertainty?

**Clinical screening.** Sleep staging with polysomnography can often be used to indicate or rule out the presence of specific sleep disorders. In Figure 6.3, we illustrate how an imperfect (as assessed by its accuracy) but well-calibrated wearable-based sleep staging model $p_\theta(y_{1:T}|\boldsymbol{X}^M_{1:T})$ could be used as an effective screening tool. If the results are conclusive, i.e. the uncertainty in the outputs of interest are low, then these could be used to inform clinical decision-making. If not, the ubiquity and ease of wearables means that multiple nights of data could be used to reduce some types of uncertainty e.g. sensor noise from loose attachment. Otherwise, the patient could be referred for alternative tests, such as a full PSG exam. Even if the outputs are only conclusive 80% of the time, this could still significantly reduce (5x) the need for more time-consuming and expensive tests.

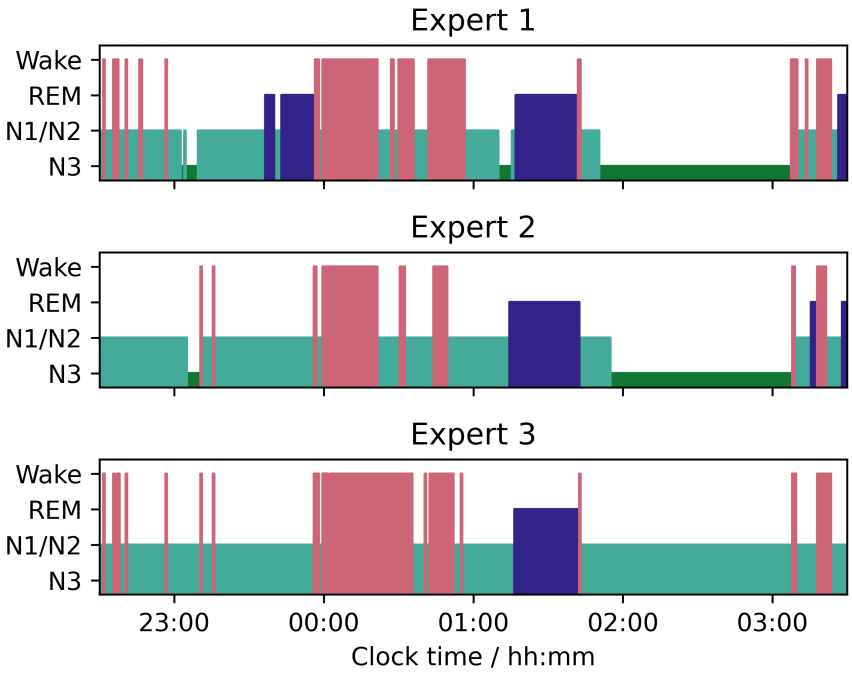



Figure 6.2. **Example of long-range inter-expert disagreement in sleep staging.** Expert 1 believed a REM cycle occurred just before 00:00. Experts 1 and 2 believed a period of N3 (deep) sleep occurred between 02:00 and 03:00.

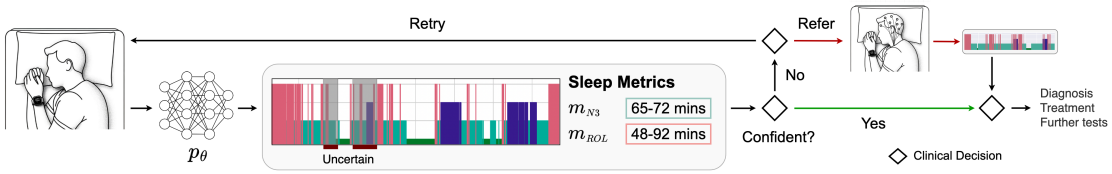



Figure 6.3. **Sleep staging under uncertainty using wearable signals.** With a *calibrated* posterior $p_\theta(y_{1:T}|\boldsymbol{X}_{1:T}^M)$, simple measures of uncertainty could be used to inform a clinician about the reliability of the outputs. For example, by providing confidence intervals on sleep metrics such as total deep sleep $m_{N3}$ and REM onset latency $m_{ROL}$.

An important consideration in the design of such a workflow is balancing the interpretability and granularity of information presented to the clinician. Alternative approaches have been proposed for presenting sleep staging uncertainty, such as hypnodensity charts [18]. However, if interpreting and processing the outputs results in significant additional work for the clinician, this will limit adoption [68].

**Evaluating treatment efficacy.** To illustrate another example of how uncertainty measures could be used in practice, suppose that we wished to investigate if treatment $I$ is effective against insomnia in a large clinical trial dataset $\mathcal{D}$. A

typical endpoint for such a trial might be whether or not administering $I$ leads to a statistically significant decrease in sleep onset latency $m_{SOL}$ (6.1) by comparing the mean values in the treatment and non-treatment arms.

$$\begin{aligned} m_{SOL} &= f_{SOL}(y_{1:T}) = t_S \\ \text{and} \quad t_S &= \min\{t \in \{1, \ldots, T\} : y_t \neq \text{Wake}\} \end{aligned} \tag{6.1}$$

After gathering the dataset $\mathcal{D}$, nights with high uncertainty in $m_{SOL}$ estimated from the algorithm could be discarded, with analysis performed over the remainder. One caveat is that for an analysis to hold, we must ensure that any uncertainty is independent of $I$. If the uncertainty depended on $I$, this filtering of uncertain nights could introduce bias, potentially altering the comparison between the two arms. However, checking for independence would be achievable with simple statistical tests.

**Other applications.** Calibrated measures of uncertainty could also be used to aid applications such as biomarker discovery or to guide the acquisition of new training data for model improvement (active learning [47]), for example, by identifying if high uncertainty is caused by conditions that are under-represented in the training data.

### 6.2.3 What's wrong with existing methods?

Most prior sleep staging methods, including those of the previous three chapters, learn a factorized posterior distribution (6.2) over the output sleep stages:

$$p_\theta(y_{1:T}|\boldsymbol{X}_{1:T}^M) = \prod_{t=1}^{T} p_\theta(y_t|\boldsymbol{X}_{1:T}^M) \tag{6.2}$$

The argmax (6.3) over the output probabilities from the model is then used to form the output sequence of sleep stage classifications:

$$\begin{aligned} \hat{y}_{1:T} &= \pi_\theta(\boldsymbol{X}_{1:T}^M) = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T\} \\ \hat{y}_t &= \arg\max_{y_t} p_\theta(y_t|\boldsymbol{X}_{1:T}^M) \end{aligned} \tag{6.3}$$

After selecting the sequence of sleep stages that maximise the posterior probabilities, the learnt probability distribution $p_\theta$ is usually discarded, giving a deterministic, learnt mapping $\pi_\theta$ from input signals to sleep stages. However, these probabilities

could instead be used to take a principled approach toward uncertainty quantification. Phan *et al.* [166] showed that the predictive entropy of an EEG-based model could be used to divide classifications into high and low-confidence sets, such that only low-confidence outputs are sent for human review, thereby reducing the manual annotation time. However, a 'human-in-the-loop' workflow [85, 101], is arguably only possible for EEG-based approaches, since there are no guidelines for scoring sleep stages from alternative signals acquired from wearables such as smartwatches. InsightSleepNet [148] similarly used an energy-based threshold to divide classifications from a PPG-based model into high and low-confidence sets, but how these could be interpreted by a clinician is not discussed.

**Posterior sampling.** In addition to providing measures of uncertainty over specific sleep stages $y_t$, a model should also be able to produce measures of uncertainty for sleep metrics, to enable the applications discussed in Section 6.2.2. The learnt posterior of a sleep staging model implicitly defines a probability distribution over sleep metrics $M_i \sim p_\theta(m_i|\boldsymbol{X}_{1:T}^M)$, which can be sampled from by first sampling sleep stages and then evaluating the resulting sleep metric:

$$Y_{1:T} \sim p_\theta(y_{1:T}|\boldsymbol{X}_{1:T}^M) \tag{6.4}$$

$$M_i = f_i(Y_{1:T}) \tag{6.5}$$

However, directly sampling output sequences $Y_{1:T}$ from a factorized posterior (6.2) ignores the temporal structure of output sleep stages, because each output is conditionally independent given the input signals $\boldsymbol{X}_{1:T}^M$. Direct sampling leads to 'spiky', physiologically implausible hypnograms that overestimate transitions between sleep stages. A factorized posterior cannot produce diverse hypnogram samples with long-range output dependencies such as those seen earlier in Figure 6.2. In turn, this means that a model that assumes a factorized posterior cannot be used to produce measures of uncertainty over sleep metrics that are sensitive to

the time-series structure, e.g. REM onset latency $m_{ROL}$ (6.6).

$$\begin{aligned} m_{ROL} &= f_{ROL}(y_{1:T}) = t_R - t_S \\ \text{where} \quad t_R &= \min\{t \in \{1, \ldots, T\} : y_t = \text{REM}\} \\ \text{and} \quad t_S &= \min\{t \in \{1, \ldots, T\} : y_t \neq \text{Wake}\} \end{aligned} \tag{6.6}$$

Only one prior work has addressed this flawed modeling assumption. van Gorp *et al.* proposed U-Flow [225], a conditional normalising flow network for EEG-based sleep staging which produces hypnogram samples conditioned on the input signals. This results in improved uncertainty quantification over sleep metrics, albeit with lower overall accuracy compared with factorized baselines.

**Quality of uncertainty.** Whilst some models such as SleepTransformer [166], U-Flow [225], and InsightSleepNet [148] have been used to produce measures of uncertainty, these measures are only estimates of the true predictive uncertainty. The reliability of these uncertainty estimates is dependent on the model itself, an issue which few works in the automated sleep staging literature have addressed. Neural networks, the predominant class of model for automated sleep staging, are known to suffer from poor calibration [72], particularly for out-of-distribution data [156]. Fiorillo *et al.* [61] have used Monte Carlo dropout and label smoothing to improve model calibration; meanwhile feature space density estimation has been used to detect out-of-distribution inputs [84].

## 6.3 Methods

### 6.3.1 SleepDecoder

To address the issues raised in Section 6.2.3, we first introduce SleepDecoder, an encoder-decoder architecture for sleep staging. An overview of the proposed model architecture is illustrated in Figure 6.4. SleepDecoder addresses the flawed factorized posterior assumption used by most existing models, enabling fully auto-regressive sampling (6.7) of sleep stage sequences during inference:

$$p_\theta(y_{1:T}|x_{1:kT}) = \prod_{i=1}^{T} p_\theta(y_i|y_{<i}, x_{1:kT}) \tag{6.7}$$

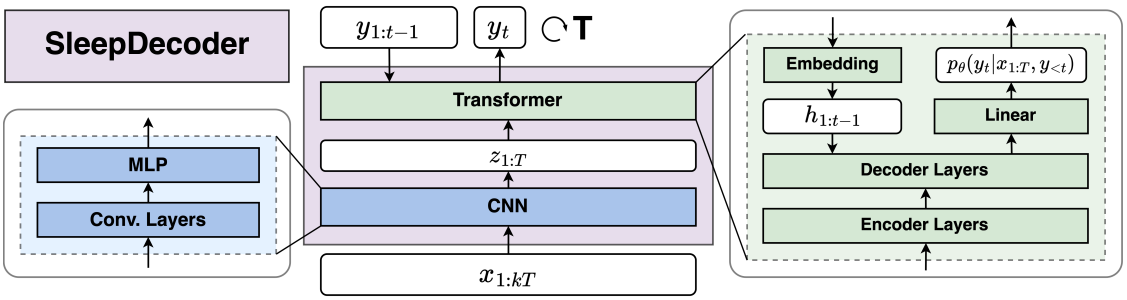


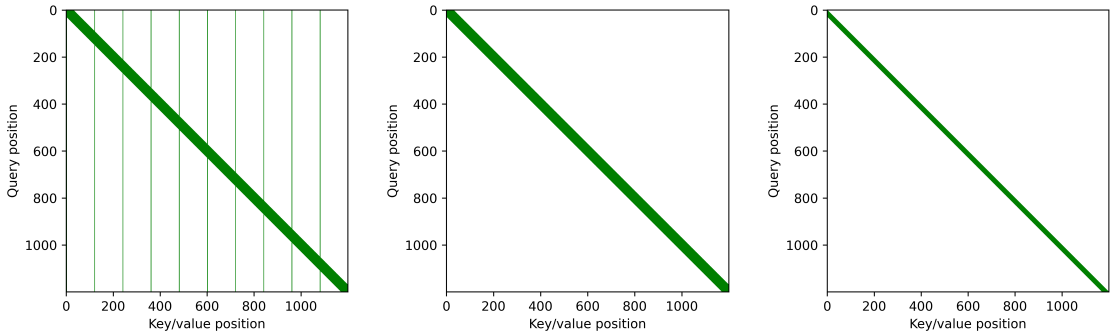

Figure 6.4. **SleepDecoder architecture.** A convolutional neural network (CNN) turns the input signal $x_{1:kT}$ into a sequence of features $z_{1:T}$. A Transformer then autoregressively predicts the sleep stage at each time-step $t$ using the feature sequence $z_{1:T}$ and a learnt embedding of the existing output sequence $h_{1:t-1}$.

**CNN Encoder.** As in the previous chapter, we use a similar CNN design to SleepPPG-Net [110] to turn the high-frequency input signal $x_{1:kT}$ with relative sampling rate $k$ into a sequence of feature vectors $z_{1:T}$. This takes 10 h (T=1200) single-channel ECG signals $x_{1:kT}$ and produces a sequence of features $z_{1:T}$. The input time-series signal is resampled to approximately 34.13 Hz such that $k = 1024$, which maintains the alignment of signals to sleep stages during the down-sampling operations in the intermediate convolutional layers.

**Transformer.** We use a Transformer [228], which takes the sequence of features $z_{1:T}$ produced by the CNN encoder and the existing output sequence $y_{1:\tau}$ as inputs and produces a conditional distribution over the next element in the output sequence $p_\theta(y_{\tau+1}|x_{1:T}, y_{1:\tau})$. As described in the original Transformer paper [228], a causal self-attention mask in the decoder portion of the Transformer allows all sequence positions $\tau \leq T$ to be decoded in parallel during training, rather than sequentially as in recurrent models such as LSTMs [86]. By default, we use $L_E = 6$ encoder layers and $L_D = 4$ decoder layers.

A learnt embedding table $\boldsymbol{E} \in \mathbb{R}^{d\times(C+1)}$ turns each discrete sleep stage $y_t$ into a feature vector $h_t \in \mathbb{R}^d$ for the Transformer. As in previous chapters, we use a feature dimension $d = 128$ throughout the model. The embedding table contains an additional learnable entry to handle unscored sleep epochs e.g. due to signal noise. We use rotary positional embeddings [208] for both the input and output feature sequences passed to the Transformer.

We use local-global attention in both the encoder and decoder layers [24]. This prevents the model from using all elements in the input and output sequences that are very far from the current output position. The masks applied to the attention matrices for the encoder self-attention, decoder self-attention and encoder-decoder cross-attention layers are visualised in Figure 6.5. The encoder self-attention mask allows each Transformer encoder layer to use a local window of 30 elements plus 10 fixed global positions. This is designed to encourage the model to focus on local information in the input sequence, with a similar inductive bias to a CNN, whilst still allowing the mixing of global information within each Transformer layer.




(a) Encoder (b) Encoder-decoder (c) Decoder

Figure 6.5. **Visualisation of attention masks used in the SleepDecoder model.** (a) Encoder layers can attend to 30 elements in each direction from the current position, plus 10 global positions. (b) Encoder-decoder layers can attend to 30 elements in each direction in the input sequence. (c) Decoder layers can attend to the previous 30 elements in the output sequence.

Hyper-parameters such as the number of encoder and decoder layers, and the context lengths for each layer were chosen using a simple grid sweep using validation set loss as the objective criterion.

### 6.3.2 Model calibration and posterior sampling

**Deep ensembles.** Rather than use the output of a single SleepDecoder model, we use a deep ensemble [117], which aggregates the outputs of multiple models trained from different initial random weights. This simple technique has proven effective for improving model predictive uncertainty over various benchmarks, particularly under dataset shift and for out-of-distribution inputs [156]. Intuitively, the inherent

randomness of neural network optimisation means that for test inputs far from the training data the different models in the ensemble should produce different outputs.

**Temperature scaling.** The 'probabilities' produced by neural network classifiers tend not to be well calibrated, i.e. they do not reflect the empirical probability of the classifier being correct, even for in-distribution data [72]. A perfectly calibrated sleep stage classifier would mean that, for example, 20% of all samples where $p_\theta(y_t = \text{Wake}\,|x_{1:kT}) = 0.2$ were actually Wake. The goal of post-hoc calibration methods such as temperature scaling is to improve the calibration of an already-trained classifier. These generally work by adjusting the output probabilities of the classifier, using another function which is learnt from the validation set.

As previously discussed in Section 2.3, the final layer of most deep neural network classifiers is a softmax function:

$$\sigma(\boldsymbol{l})^{(c)} = \frac{\exp(l^{(c)})}{\sum_{c=1}^{C} \exp(l^{(c)})} \tag{6.8}$$

which converts an intermediate vector $\boldsymbol{l} \in \mathbb{R}^C$, often referred to as the 'logits', into a valid categorical distribution over the $C$ output classes.

Temperature scaling re-calibrates the softmax layer by scaling the vector of logits by a scalar temperature T, i.e. $\sigma(\boldsymbol{l}) \to \sigma(\boldsymbol{l}/T)$. As $T \to 0$, the output distribution tends toward a point mass on the most confident class. As $T \to \infty$, the output distribution tends toward a uniform distribution over the classes. After training, i.e. finding neural network parameters that minimise the negative log-likelihood of the training set, temperature scaling finds the optimal $T$ that minimises the negative log-likelihood of the validation set. This simple step has been shown to improve the calibration of neural network output probabilities [72].

**Temperature-scaled deep ensembles.** To use temperature scaling with deep ensembles, we followed the 'pool-then-calibrate' approach of Rahaman and Thiery [175], averaging the logit vectors produced by each encoder-decoder model in the

ensemble before performing temperature scaling, i.e.:

$$p_\theta(y_t|y_{<t}, x_{1:kT}) = \sigma\left(\frac{\tilde{\boldsymbol{l}}_t}{T}\right) \tag{6.9}$$

$$\tilde{\boldsymbol{l}}_t = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{l}_t^{(m)} \tag{6.10}$$

$$\boldsymbol{l}_t^{(m)} = f_{\theta^{(m)}}(y_{<t}, x_{1:kT}) \tag{6.11}$$

where $\theta = \{\theta^{(1)}, \ldots, \theta^{(M)}\}$ collectively denotes the parameters of the $M$ models in the ensemble and $f_{\theta^{(m)}}$ denotes the function learnt by each model. To find the optimal temperature $T$ we generated logits for the validation set, then optimised the temperature using the L-BFGS optimization algorithm [35]. This step takes under a minute to run, meaning that the computation is insignificant compared with the training of each neural network.

**Posterior sampling.** For each night of data, we drew $S = 128$ samples from the model posterior (6.7), which can be performed in parallel on a GPU. This fixed value of $S$ was chosen as an optimal trade-off between computational efficiency and performance. The notation $y_{1:T}^s$ is used to refer to sample $s$ from the model posterior. When evaluating the overall performance of SleepDecoder against expert-annotated sleep stages, the mode over posterior samples was used as the output classification for each time-step $t$, which is equivalent to the argmax of the (intractable) marginal distribution $p_\theta(y_t|x_{1:kT})$:

$$\hat{y}_t = \text{mode}(y_t^1, \ldots, y_t^S) \tag{6.12}$$

$$= \arg\max_{y_t} p_\theta(y_t|x_{1:kT}) \quad \text{as} \quad S \to \infty \tag{6.13}$$

### 6.3.3 Measures of uncertainty

After training, we wish to use the model to estimate sleep stage sequences $y_{1:T}$, sleep metrics such as $m_{REM}$ (6.21), and to produce measures of uncertainty over these outputs. Here we define simple, principled measures of uncertainty over individual sleep epochs (30-second windows), over entire nights, and over arbitrary sleep metrics $m_i$. We re-iterate that these measures are only *estimates* of the true predictive uncertainty. As we later show, the reliability of these uncertainty

estimates is heavily dependent on the model specification, which techniques like deep ensembling and temperature scaling used in this chapter aim to address.

**Epoch-level uncertainty.** For a calibrated model, when a greater number of samples from the output posterior $y_t^s$ agree, there is a greater likelihood that the modal classification (6.12) is correct. Therefore we used the frequency of the modal class (6.14) as a simple confidence measure $c_t \in [0, 1]$ for each sleep epoch. Like the mode $\hat{y}_t$, this also has a probabilistic interpretation. It is equivalent to the maximum of the marginal distribution $p_\theta(y_t|x_{1:T})$:

$$c_t = \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}(y_t^s = \hat{y}_t) \tag{6.14}$$

$$= \max_{y_t} p_\theta(y_t|x_{1:T}) \quad \text{as} \quad S \to \infty \tag{6.15}$$

**Night-level uncertainty.** To provide simple summary measures of confidence over an entire night of data, we use the average confidence:

$$\bar{c} = \frac{1}{T}\sum_{t=1}^{T} c_t \tag{6.16}$$

and the proportion of epochs with a confidence greater than some threshold $\delta$:

$$\nu_\delta = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}(c_t > \delta) \tag{6.17}$$

As we later show, this confidence threshold $\delta$ may be chosen to ensure an expected level of accuracy over epochs marked as 'confident'.

**Sleep metric uncertainty.** As discussed in Section 6.2.3, the learnt model posterior also implicitly defines a probability distribution over sleep metrics $p_\theta(m_i)$. Again assuming that $p_\theta(m_i)$ is calibrated, then for an arbitrary estimate $\hat{m}_i$ the expected squared difference can be approximated from the $S$ samples:

$$\sigma_i^2 = \frac{1}{S}\sum_{s=1}^{S}(m_i^s - \hat{m}_i)^2 \tag{6.18}$$

$$\lim_{S\to\infty} \sigma_i^2 = \mathbb{E}_{p_\theta}[(m_i - \hat{m}_i)^2] \tag{6.19}$$

i.e. $\sigma_i$ is an approximation of the root-mean-squared-difference (RMSD) between our estimate $\hat{m}_i$ and the true sleep metric $m_i$. In later experiments, we use the median over samples $\hat{m}_i = \text{median}(m_i^1, \ldots, m_i^S)$ as our sleep metric estimate and

$\sigma_i$ as a measure of uncertainty over this estimate. In addition to $m_{SOL}$ (6.1) and $m_{ROL}$ (6.6) defined in Section 6.2, we evaluate model performance for total deep sleep time $m_{N3}$ (6.20) and total REM time $m_{REM}$ (6.21).

$$m_{N3} = f_{N3}(y_{1:T}) = \sum_{t=1}^{T} \mathbf{1}(y_t = \text{N3}) \tag{6.20}$$

$$m_{REM} = f_{REM}(y_{1:T}) = \sum_{t=1}^{T} \mathbf{1}(y_t = \text{REM}) \tag{6.21}$$

### 6.3.4 Training and evaluation procedure

As in the previous chapter, all models were trained using the SHHS, MESA, MROS, WSC, CFS, CHAT and CCSHS datasets, using the same dataset splits. Again, we also report performance using the demographically balanced test set proposed by Jones *et al.* [97], referred to in this thesis as the 'Census' test set.

In addition to these datasets, we also use the Dreem Open datasets (DOD, [71]) to evaluate the SleepDecoder model. The DOD datasets consist of a 25-night dataset (DOD-H) recorded from healthy volunteers, and a 55-night dataset (DOD-O) recorded from patients diagnosed with obstructive sleep apnoea. Unlike the other datasets, each overnight PSG recording from the Dreem datasets has been annotated by five independent sleep experts. The multiple annotations available in the Dreem datasets later enable us to compare model-expert and expert-expert agreement statistics.

We use the MROS and DOD-H/DOD-O datasets as completely held-out test sets, with no data from any of these datasets used during the development of the model. For the DOD-H and DOD-O datasets, we evaluate sleep staging performance against the majority vote (mode) over the five expert annotations for each time step. In line with previous chapters, we report the overall performance of models in terms of the total Cohen's $\kappa$ statistic and model-expert agreement (accuracy) over sleep epochs.

All models were trained to minimize the cross-entropy loss between the model output posterior $p_\theta(y_{1:T}|\boldsymbol{x}_{1:kT})$ and expert-annotated sleep stages $y^e_{1:T}$ using the AdamW optimizer [134] and the same optimisation hyper-parameters as the previous chapter (see Section 5.4.2).

## 6.4 Experiments

In this section, we evaluate the performance of the SleepDecoder model, including the effect of using deep ensembles and temperature scaling, compared with encoder-only variants. For brevity, we often refer to the best-performing variant of SleepDecoder, which uses temperature scaling and deep ensembling, as simply a 'decoder ensemble' throughout this section.

### 6.4.1 Overall performance

In Table 6.1, we compare the performance of the SleepDecoder model with two encoder-only ($L_D = 0$) variants, one with the same number of encoder layers ($L_E = 6$), and another with the same number of total transformer layers ($L_E = 10$). We x

Table 6.1. Performance ($\kappa_T$) of different SleepDecoder variants.

| | | | Dataset | | | | |
|---|---|---|---|---|---|---|---|
| $L_E$ | $L_D$ | DE | SHHS | Census | MROS | DOD-H | DOD-O |
| 10 | 0 | ✗ | 0.716 | 0.778 | 0.717 | 0.693 | 0.691 |
| | | ✓ | 0.737 | 0.797 | 0.743 | 0.724 | 0.724 |
| 6 | 0 | ✗ | 0.720 | 0.781 | 0.718 | 0.718 | 0.693 |
| | | ✓ | 0.740 | **0.799** | 0.744 | 0.741 | 0.720 |
| 6 | 4 | ✗ | 0.736 | 0.775 | 0.735 | 0.747 | 0.730 |
| | | ✓ | **0.747** | 0.784 | **0.751** | **0.756** | **0.743** |

$L_E$ - Encoder layers. $L_D$ - Decoder layers. DE - Deep ensemble.

With the exception of the Census dataset, the SleepDecoder model, which uses both encoder and decoder layers, performs best. Using deep ensembles leads to improved performance across all model variants and datasets. Whilst it is encouraging that the auto-regressive SleepDecoder model generally performs better than encoder-only baselines, and similarly to the best-performing models of Chapter 5, overall classification performance is not the focus of this chapter. Instead, we are more interested in the quality of predictive uncertainty produced by the model, which we begin evaluating in the next subsection.

### 6.4.2 Uncertainty over sleep stages

**Calibration.** Figure 6.6 shows 'reliability diagrams' [150] for the SHHS test set for different model variants. These plots show the fraction of correct classifications plotted against the epoch confidence $c_t$. To produce a reliability diagram, model classifications are sorted by confidence and divided into percentile bins. The x- and y-coordinates of each point on the reliability diagram are then given by the average confidence within each bin and the fraction of correct classifications within each bin respectively. For a perfectly calibrated model, this should yield a straight line. In Figure 6.6 we see that an encoder-decoder architecture naturally leads to improved calibration, which is further improved by temperature scaling.

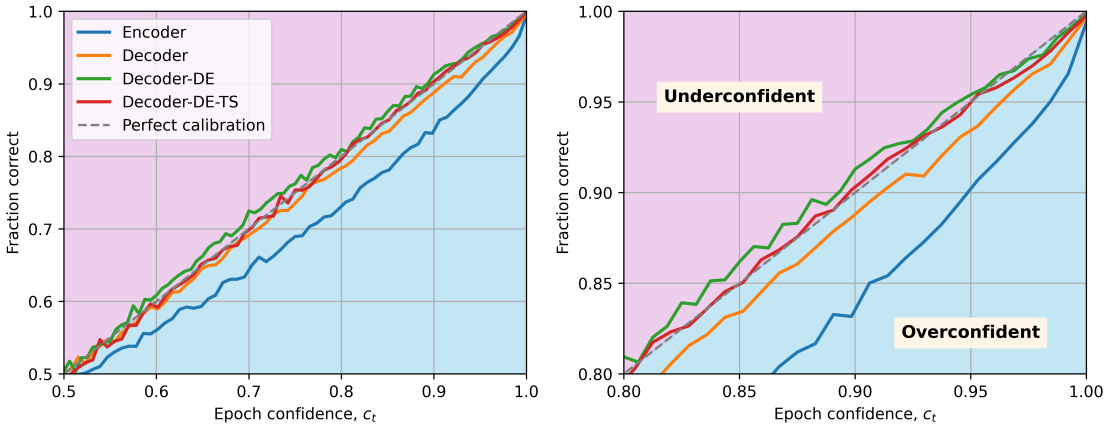



Figure 6.6. **SleepDecoder calibration curves on the SHHS test set.** (left) $c_t \in [0.5, 1]$. (right) Zoomed, $c_t \in [0.8, 1]$. Calibration curves are plotted for encoder-only, encoder-decoder, deep ensemble (DE) and temperature scaled (TS) model variants.

**Epoch-level uncertainty.** Using the held-out MROS test set, Figure 6.7 shows the Cohen's $\kappa$ and accuracy over confident epochs plotted against the proportion of epochs marked as confident $\nu_\delta$, as we sweep over a minimum threshold $\delta$ on our epoch-level confidence measure (6.14) i.e. $c_t \geq \delta$ for $\delta \in [0, 1]$. By labelling 10% of epochs as uncertain, we see that the accuracy of our best-performing model increases from around 84% to 88% and the Cohen's $\kappa$ from around 0.75 to over 0.80 over the remaining epochs. Beyond solely improving raw accuracy statistics, improving model calibration and uncertainty quantification can improve

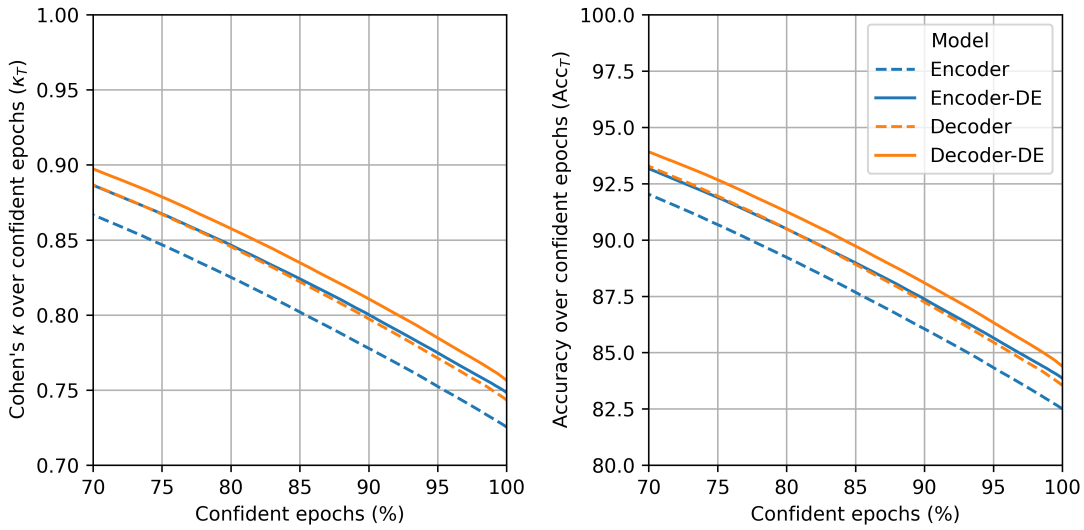
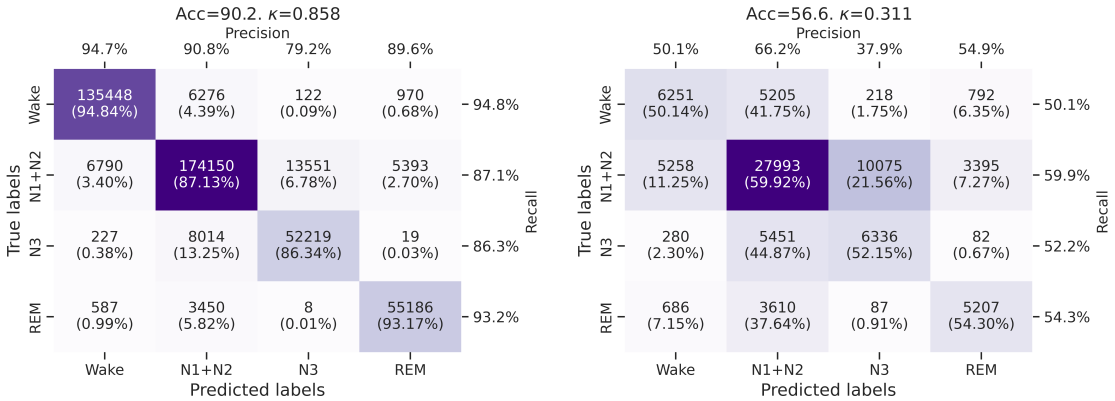



Figure 6.7. **Performance-confidence frontiers for the MROS test set.** Model performance (left - Cohen's $\kappa$, right - accuracy) calculated over confident epochs is plotted against the proportion of confident epochs as we vary the epoch confidence threshold $c_t \geq \delta$ for $\delta \in [0, 1]$.

this two-dimensional view of performance, improving the utility of the model for the workflow illustrated in Figure 6.3.

**When is the model uncertain?** Using a fixed confidence threshold $\delta = 0.65$, Figure 6.8 shows separate confusion matrices for confident and unconfident epochs over the Census test set. Here we can observe the sources and relative frequencies of misclassification of the different sleep stages, with most disagreement occurring between Light (N1+N2) and Deep (N3) sleep.




(a) Confident epochs, $c_t \geq 0.65$.

(b) Unconfident epochs, $c_t < 0.65$.

Figure 6.8. Confusion matrices for confident and unconfident epochs on the Census test set using a fixed confidence threshold.

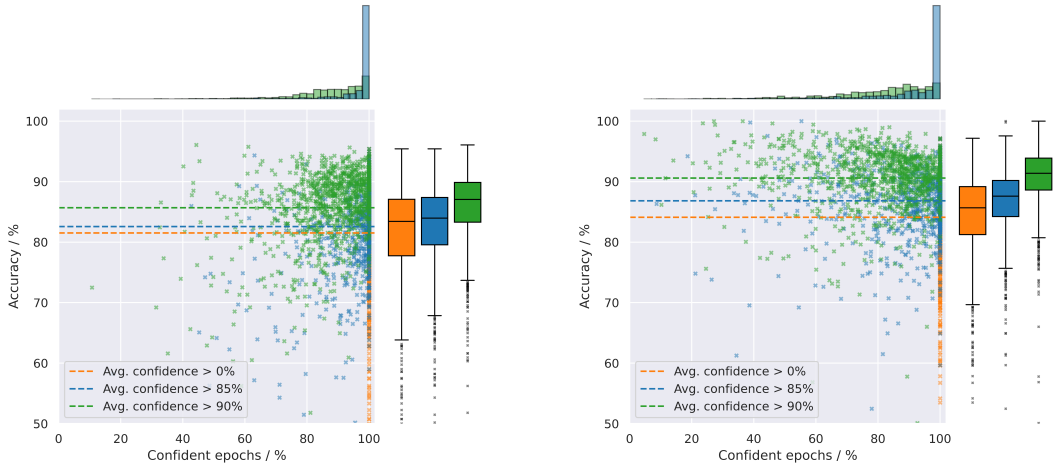
**Night-level uncertainty.** With a calibrated posterior, the expected accuracy is equal to the average confidence $\bar{c}$, which allows us to take a principled approach to uncertainty. For example, to achieve an expected level of accuracy over sleep epochs marked as 'confident', we can set the epoch confidence threshold $c_t > \delta$ such that the average confidence over confident epochs is greater than or equal to the desired accuracy.

Summary results of using this technique are shown in Figure 6.9, which plots the average accuracy over confident epochs against the percentage of confident epochs for individual nights from the MROS test set. Ignoring measures of uncertainty, and designating all epochs as 'confident', the average accuracy is around 83% (orange line). As we set stricter thresholds on the minimum average confidence (85%, 90%) for each night, the number of epochs marked as confident decreases and the average accuracy increases. For the temperature-scaled decoder ensemble (Figure 6.9b), this does indeed result in an expected accuracy equal to the average confidence (e.g. dashed green line), indicating that the model posterior remains well-calibrated, even on the held-out MROS test set.

**Example nights.** Figures 6.10 and 6.11 illustrate the effect of marking sleep epochs as uncertain for two example nights of data, one with few ($< 20\%$) uncertain epochs and another with many ($> 50\%$) uncertain epochs. In each case, the lowest confidence epochs are marked as uncertain (shaded grey) such that the confidence over the remaining epochs exceeds 90%. The empirical marginal distribution $p_\theta(y_t|x_{1:T})$ for each sleep stage is also visualised using a hypnodensity plot [18].

For the night with few uncertain epochs in Figure 6.10, we observe that most of the uncertainty observed throughout the night is between Light and Deep sleep, with all REM cycles correctly identified by the model and very high overall agreement over confident epochs ($\kappa_c = 0.834$).

Meanwhile, for the night with high uncertainty (Figure 6.11), the hypnodensity plot shows that there is uncertainty between light and deep sleep earlier in the night, whilst toward the end of the night (beyond 6 hours elapsed) there is a long period of uncertainty between multiple sleep stages. Investigating the underlying ECG




(a) Unscaled encoder-only model. (b) Temperature-scaled decoder ensemble.

Figure 6.9. **Calibrated uncertainty enables an expected level of accuracy by marking epochs as uncertain.** In each subfigure: (top) Histogram of confident epochs per night for different confidence thresholds. (centre) Per-night accuracy over confident epochs plotted against the proportion of confident epochs, for different confidence thresholds $\delta$. (right) Box plots of per-night accuracies over confident epochs. Results were calculated using nights from the held-out MROS test set. In (b), we observe that requiring the average confidence to be at least 90% results in an average accuracy of around 90%.

signal for this night of data, around the 6-hour mark there is noticeable additional baseline wander of the ECG signal. This is likely the cause of the high uncertainty observed, making it more difficult for the model to extract physiological information under increased noise. Despite high overall uncertainty, there is low uncertainty in the timing of the first REM cycle, enabling REM onset latency to be estimated accurately. This highlights how certain metrics of interest may still be extractable with high confidence even when there is high overall uncertainty.

### 6.4.3 Uncertainty over sleep metrics

**Posterior samples.** The top row of Figure 6.12 shows hypnograms produced for the same night of data by three different experts from a recording in the DOD-O dataset. In the middle and bottom rows, for each expert hypnogram, we plot the closest[3] sample hypnogram from the output posterior $p_\theta(y_{1:T}|\boldsymbol{x}_{1:kT})$ for the

[3]As measured by Cohen's $\kappa$ statistic.

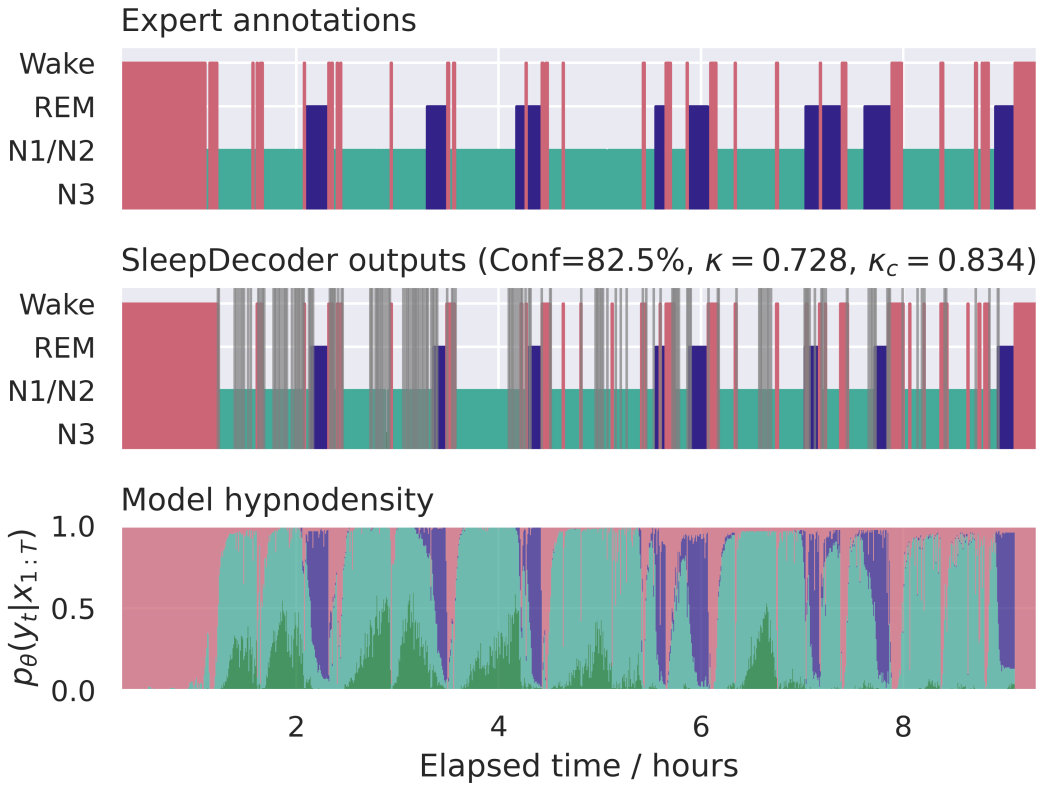



Figure 6.10. **Example night with low uncertainty from the MROS test set**. (top) Expert-annotated sleep stages. (middle) SleepDecoder outputs, unconfident sleep epochs where $c_t < \delta$ are shaded gray. $\kappa_c$ and $\kappa$ denote the Cohen's $\kappa$ over confident epochs and over all epochs respectively. (bottom) Hypnodensity plot showing the empirical marginal distribution $p_\theta(y_t|x_{1:T})$.

SleepDecoder model and SleepPPG-Net [110], a sleep staging model that employs a factorised posterior. Here we see that the auto-regressive formulation of the SleepDecoder model enables it to produce hypnogram samples that more closely reflect the diversity observed among human experts. Meanwhile, directly sampling from a factorised posterior leads to 'spiky', physiologically implausible hypnograms.

**Visual comparison with experts.** In Figure 6.13, we plot model-derived and expert-derived sleep metrics across all recordings from the multi-scored DOD datasets, with nights ordered by the median expert-derived sleep metric. For most nights of data, we see that the model-derived estimates generally fall within the range of expert-derived values. However, there is frequent disagreement between expert values, particularly in the amount of deep sleep $m_{N3}$ (see Figure 6.13a).

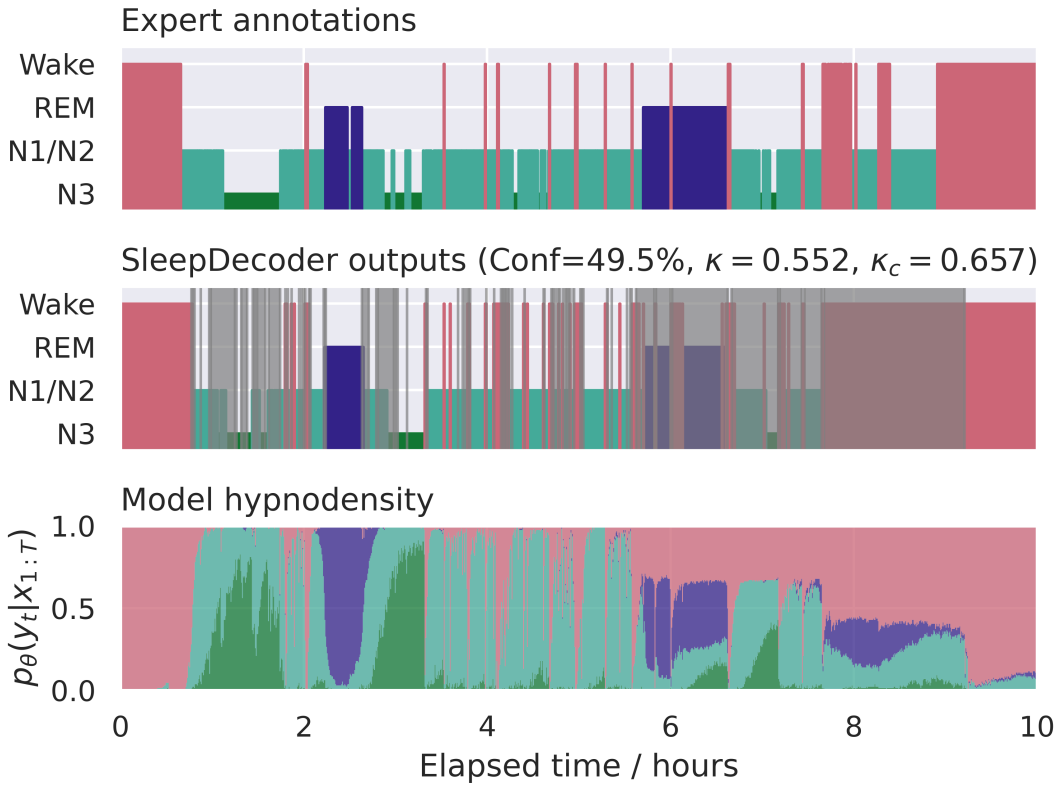



Figure 6.11. **Example of a night with high uncertainty from the MROS test set**. After 6 hours, there is a noticeable additional baseline wander of the ECG. (top) Expert-annotated sleep stages. (middle) SleepDecoder outputs, unconfident sleep epochs are shaded gray. $\kappa_c$ and $\kappa$ denote the Cohen's $\kappa$ over confident epochs and over all epochs respectively. (bottom) Hypnodensity plot showing the empirical marginal distribution $p_\theta(y_t|x_{1:T})$ for each sleep stage.

**Sleep metric confidence.** Figure 6.14 shows the difference between model-derived and expert sleep metrics, $e_i$, plotted against the sleep metric confidence measure $\sigma_i$. Since $\sigma_i$ is the expected RMSD between the estimated and true sleep metric (6.19), if $p_\theta$ is well-calibrated, the variance of $e_i$ should increase with $\sigma_i$. Lines indicating $e_i = \pm\sigma_i$ and $e_i = \pm 1.96\sigma_i$ are also shown to aid this visualisation. Here we see that, when using a decoder ensemble, $\sigma_i$ provides an informative measure of uncertainty over sleep metrics. For example, comparing Figure 6.14a and Figure 6.14b, $\sigma_{N3}$ is generally much higher than $\sigma_{REM}$ for each night of data, in line with the increased variance in $e_{N3}$. In contrast, the factorised posterior of an encoder-only model means that $p_\theta(m_i)$ is a poor approximation, which means that $\sigma_i$ is not predictive of the true variation in $e_i$.

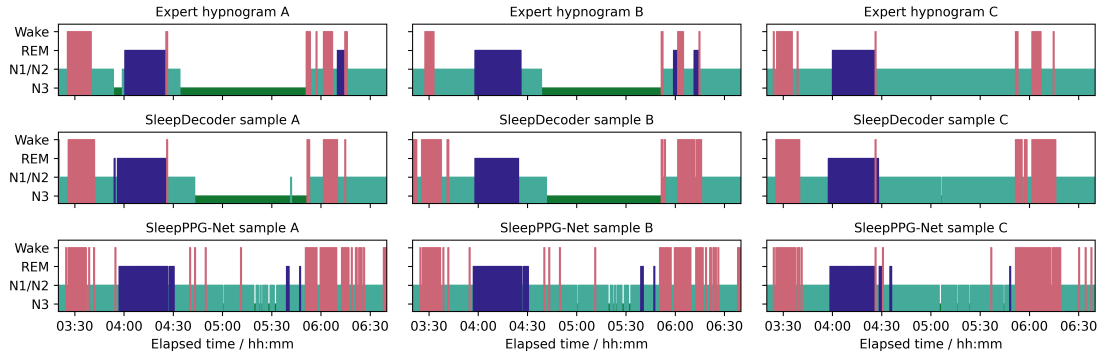
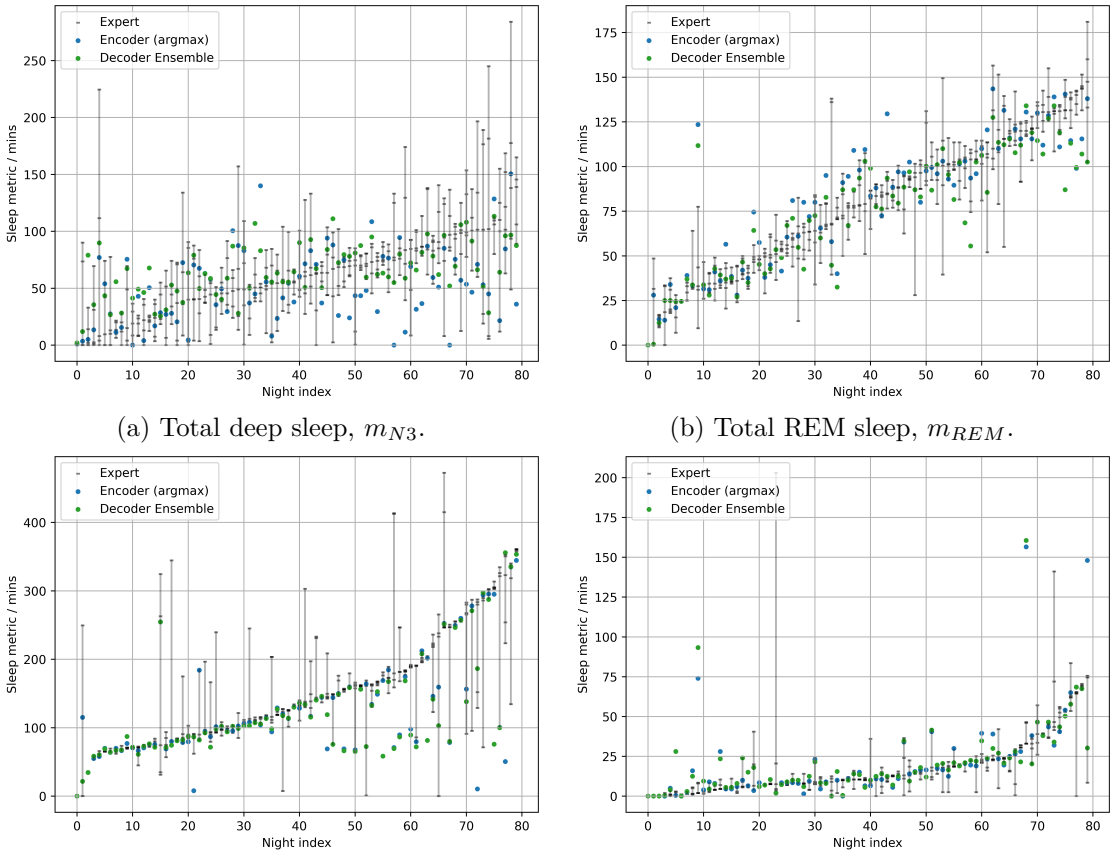



Figure 6.12. **SleepDecoder models long-range uncertainty similarly to human experts.** Example of expert-annotated hypnograms and posterior samples from $p_\theta(y_{1:T}|\boldsymbol{x}_{1:kT})$ for the same night of data taken from the DOD-O dataset. (top row) Expert-annotated sleep hypnograms using the PSG recording. (middle row) The closest posterior sample to each expert hypnogram using SleepDecoder applied to the ECG. (bottom row) The closest posterior sample to each expert hypnogram using SleepPPG-Net [110] applied to the ECG. Closeness is evaluated using Cohen's $\kappa$ statistic.



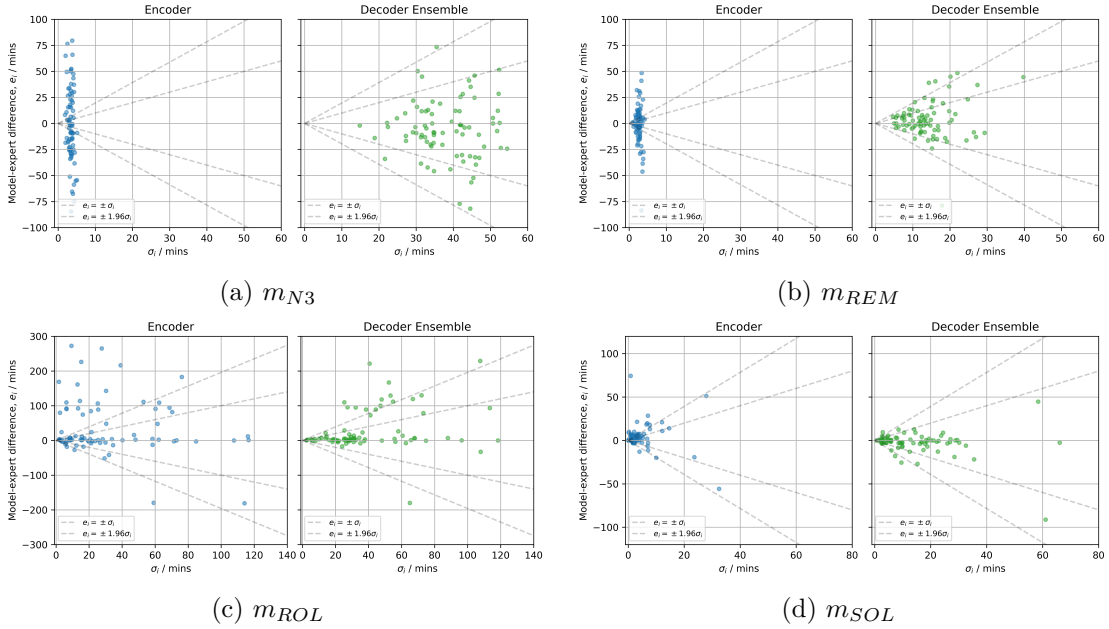

(a) Total deep sleep, $m_{N3}$.

(b) Total REM sleep, $m_{REM}$.

(c) REM onset latency, $m_{ROL}$.

(d) Sleep onset latency, $m_{SOL}$.

Figure 6.13. Model-derived and expert-derived sleep metrics for recordings from the DOD datasets. Nights are ordered by the expert-derived median value. Vertical bars indicate the range of expert-derived values.




(a) $m_{N3}$ (b) $m_{REM}$

(c) $m_{ROL}$ (d) $m_{SOL}$

Figure 6.14. **Model-expert sleep metric differences $e_i$ plotted against $\sigma_i$ for recordings from the DOD datasets.** Errors are calculated against the median expert value. Using a decoder ensemble, $\sigma_i$ is predictive of the variation in model-expert differences.

Generally, Figure 6.14 indicates that the SleepDecoder model can produce sleep metric estimates with low bias and that it is possible to know when the model is less likely to be accurate, i.e. when there is higher variance in $e_i$. For applications such as evaluating treatment efficacy discussed in Section 6.2.2, overnight recordings could be discarded where $\sigma_i$ is above a predetermined threshold chosen using a validation set prior to deployment. For example, from Figure 6.14d, a threshold $\sigma_{SOL} < 20$ could be a reasonable choice to exclude outliers similar to those observed.

Whilst these results are encouraging, the predictive uncertainty of the SleepDecoder model is not perfect. For example, there are instances where $\sigma_i$ greatly underestimates the true difference $e_i$, particularly for many estimates of $m_{ROL}$ in Figure 6.14c. However, one limitation of this analysis is that we have used the median expert value as a proxy for the true sleep metric $m_i$. This means that some instances where $\sigma_i$ underestimates $e_i$ could be due to inaccuracies in the expert-derived values, rather than due to inaccuracies in the model's estimate of

uncertainty.[4] Given the high variance in some of the expert-derived metrics observed, acquiring additional expert annotations for nights of data with high inter-expert disagreement would help to improve the robustness of this evaluation.

### 6.4.4 Out-of-distribution detection

Next, we evaluated the ability of the model to handle corrupted, out-of-distribution inputs, using a corrupted version of the Dreem datasets. For each recording from the DOD datasets, we applied one of four possible corruptions to the ECG signals (chosen uniformly at random):

1. Baseline wander with a high-amplitude sinusoid of frequency 0.1 Hz.
2. Low Gaussian noise ($\sigma = 1$).
3. High Gaussian noise ($\sigma = 3$).
4. $N$ random gaps of length $T$ minutes, where $N \sim \mathcal{U}(20, 40)$ and $T \sim \mathcal{U}(0, 15)$.

An example of each corruption/augmentation is shown in Figure 6.15. These corrupted versions were then used to form a new out-of-distribution dataset which we refer to as DOD-Augmented.

Figure 6.16 plots model accuracy over confident epochs against the proportion of confident epochs $\nu_\delta$ for each night of data for both the DOD and DOD-Augmented datasets, using a fixed epoch confidence threshold $c_t > \delta = 0.65$. We plot results for encoder, decoder and decoder ensemble model variants. We apply temperature scaling to all model variants. With a decoder ensemble, this simple confidence threshold gives a clear separation of out-of-distribution DOD-Augmented recordings, resulting in most nights having fewer than 20% of epochs marked as confident. Empirically, these results were robust to the choice of $\delta$, with values between 0.55 and 0.7 giving similarly clear separations between the two datasets. In practice, when there is such substantial uncertainty over a night, it may be better to discard the entire night of data, particularly given the higher variance in accuracy observed.

[4]Sources of uncertainty for the model and experts are different, e.g. ECG vs. EEG signal noise.

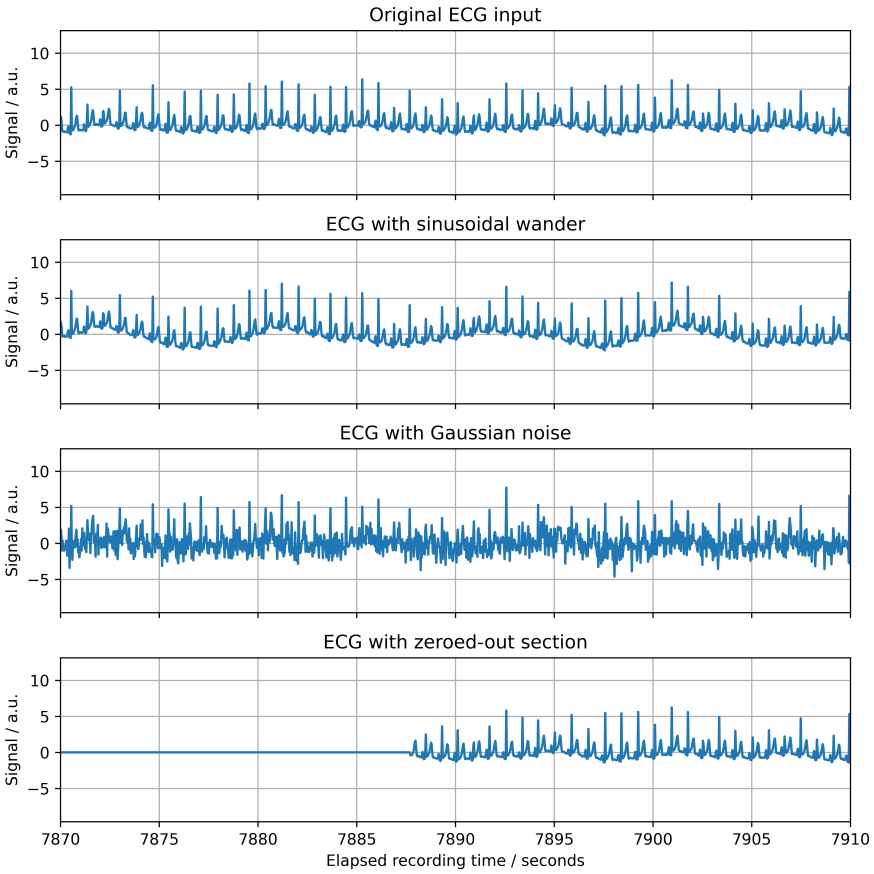
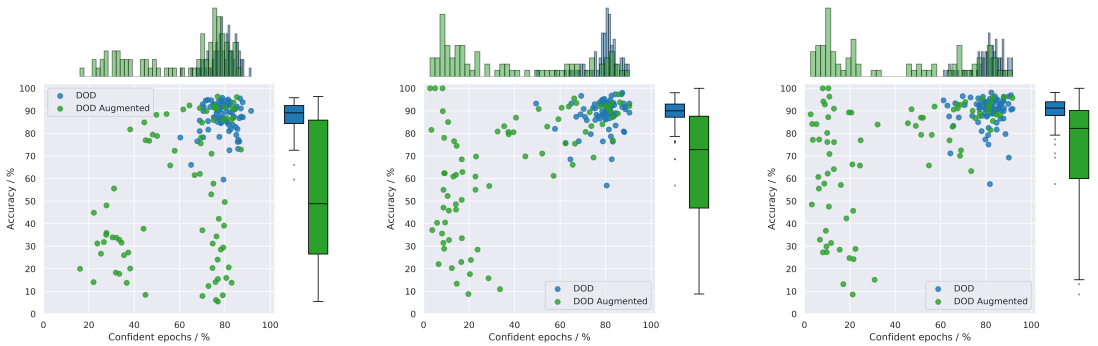



Figure 6.15. Example ECG augmentations used to form the DOD-Augmented dataset.




(a) Encoder. (b) Decoder. (c) Decoder ensemble.

Figure 6.16. **Decoder ensembles improve out-of-distribution uncertainty quantification.** Using a decoder ensemble results in a lower proportion of incorrectly classified, out-of-distribution sleep epochs being assigned a high confidence. In each subfigure: (top) Histogram of confident epochs per night for DOD and DOD-Augmented datasets. (centre) Per-night accuracy over confident epochs plotted against the proportion of confident epochs, for a confidence threshold of $c_t > 0.65$. (right) Box plots of per-night accuracies over confident epochs.

### 6.4.5 Comparing model-expert and inter-expert agreement

In this section, we compare the model-expert agreement of our best-performing model with the agreement observed between human experts. Jones *et al.* [97] claimed that their ECG-based sleep staging model achieves 'expert-level' sleep staging accuracy. However, this claim is made by comparing the performance of their model on the Census test set with inter-expert agreement statistics from a meta-analysis using different datasets. Inter-expert agreement is influenced by factors such as the prevalence of sleep disorders and the quality of the recording equipment [224], so directly comparing Cohen's $\kappa$ statistics across different datasets may be inappropriate. In this section, we use the multi-scored DOD datasets to directly compare model-expert and inter-expert agreement over the same sets.

For each night of data, we calculate the Cohen's $\kappa$ between the model output and the scorer majority vote labels, denoted $\kappa^{\theta}$, and the 5 Cohen's $\kappa$ agreement values between each scorer $e$ and the majority vote labels, each denoted $\kappa^{e}$. We then perform a series of Wilcoxon signed-rank tests to establish the superiority of the model or experts for various model confidence thresholds $\delta$. Summary results for these experiments are provided in Table 6.2.

Table 6.2. **Model-expert agreement statistics and expert-expert agreement statistics for the DOD datasets.** Model performance is calculated over confident epochs for varying confidence thresholds $\delta$. Cohen's $\kappa$ values are calculated against the expert majority vote for each recording. Farthest and nearest refer to the lowest and highest expert-majority agreement for each recording. $q_n$ refers to the $n^{\text{th}}$ percentile value over recordings.

| | | Cohen's $\kappa$ | | | Confident epochs, $\nu_\delta$ / % | | |
|---|---|---|---|---|---|---|---|
| | $\delta$ | $q_{25}$ | $q_{50}$ | $q_{75}$ | $q_{25}$ | $q_{50}$ | $p_{75}$ |
| SleepDecoder | 0.00 | 0.720 | 0.788 | 0.842 | 100.0 | 100.0 | 100.0 |
| | 0.55 | 0.741 | 0.810 | 0.840 | 90.4 | 93.1 | 94.9 |
| | 0.65 | 0.814 | 0.854 | 0.896 | 78.6 | 82.0 | 85.1 |
| | 0.70 | 0.836 | 0.884 | 0.918 | 72.3 | 76.8 | 80.3 |
| Farthest, $\min_e \kappa^e$ | - | 0.656 | 0.712 | 0.809 | - | - | - |
| Average, $\text{mean}(\kappa^e)$ | - | 0.824 | 0.851 | 0.895 | - | - | - |
| Nearest, $\max_e \kappa^e$ | - | 0.887 | 0.910 | 0.947 | - | - | - |

Firstly, without accounting for uncertainty, we find that our model performs better than the 'farthest' human scorer on average, i.e. $\kappa^{\theta} > \min_e \kappa^e$ over recordings from the Dreem datasets (0.788 vs. 0.712, $p = 0.003$). However, $\kappa^{\theta} < \text{mean}(\kappa^e)$ (0.788 vs. 0.851, $p < 1e^{-6}$), i.e. the model performs worse than the average expert.

Next, we investigate how uncertainty quantification affects these results, by varying the minimum epoch confidence $c_t$. With a confidence threshold $c_t > 0.65$, the model performs slightly better than the average expert (on average) but not to a level of statistical significance (0.854 vs. 0.851, $p = 0.25$). With $c_t > 0.7$, the model performs better than the average expert with statistical significance (0.884, vs 0.51, $p = 0.0001$).

Using a fixed confidence threshold $c_t > 0.65$ results in a median value of 82.0% of epochs marked as 'confident'. Meanwhile, Figure 6.10 previously showed that important insights can often be derived from a night when around 80% of epochs are marked as confident. This suggests that the performance of these algorithms is sufficient to derive clinically useful sleep insights reliably. Future work should assess whether these algorithms can replicate the same clinical endpoints as traditional polysomnography within clinical trials, potentially paving the way toward using less obtrusive modalities such as ECG signals as the primary input source for the validation of new treatments or for the discovery of sleep biomarkers.

## 6.5 Conclusions

This chapter has highlighted the importance of uncertainty quantification for automated sleep staging using modalities such as the ECG, and described how uncertainty could be handled within applications such as sleep disorder screening or evaluations of treatment efficacy.

Toward this goal, we introduced SleepDecoder, an encoder-decoder ensemble model for automated sleep staging which addresses the factorized posterior assumption employed by most methods described in the literature. This model was used to derive simple, principled measures of uncertainty for individual sleep epochs, entire nights, and estimates of sleep metrics. Unlike encoder-only baseline methods,

these measures of uncertainty were shown to be well-calibrated, even for completely held-out test sets, and for out-of-distribution ECG recordings augmented with corrupted segments.

Moreover, this improved uncertainty quantification was achieved without sacrificing predictive performance, with the SleepDecoder model outperforming baseline methods, and performing on a par with the best-performing approach of Chapter 5, according to the commonly used Cohen's $\kappa$ statistic. Comparing model-expert and expert-expert agreement statistics, the SleepDecoder model was able to achieve comparable sleep staging performance to human experts. By labelling fewer than 20% of sleep epochs as uncertain, the model achieved similar Cohen's $\kappa$ agreement values to an average human expert when compared with an expert consensus. Meanwhile, earlier analysis in this chapter showed that labelling a similar percentage of epochs as uncertain could still allow the derivation of important insights such as sleep and REM onset latency, depending on when the uncertain periods occurred during the recording.

This chapter has focused on single-channel ECG-based sleep staging. However, the techniques introduced in this chapter are fully compatible with the wav2sleep approach introduced in Chapter 5. These could be used to improve the performance of the model even further and to apply it to a wider range of modalities.

**Limitations.** Both an encoder-decoder architecture and deep ensembling add significant computational overhead. Deep ensembling increases the total computation performed during both training and inference by a factor of $M$, where $M$ is the size of the ensemble. Meanwhile, since the computational complexity of the attention mechanism is quadratic in the sequence length $T$, and auto-regressive sampling requires the sequential generation of the full sequence $T = 1200$, the total computational complexity of posterior sampling from an encoder-decoder model is $O(T^3)$. Running a full encoder-decoder ensemble over a single night of data with $S = 128$ hypnogram samples required approximately one minute of processing time using an NVIDIA V100 GPU, though this code has not been optimised. Because our attention matrices are sparse (see Figure 6.5), the computational complexity could

be substantially reduced from $O(T^3)$ to $O(L^3)$, where $L$ is the local context length, e.g. $L = 30$ for the decoder. In practice, since the model only needs to be run once for a single night of data during inference, this was not an issue for this research.

# 7 Conclusions

## Contents


**7.1 Review of Contributions . . . . . 141**
**7.2 Future Work . . . . . 143**
7.2.1 Vision-based sleep monitoring . . . . . 143
7.2.2 Physiological representation learning . . . . . 144
7.2.3 Clinical validation . . . . . 145
**7.3 Concluding Remarks . . . . . 146**


## 7.1 Review of Contributions

This thesis aimed to address the shortcomings in traditional sleep monitoring from manual polysomnography, by designing models and algorithms that could automatically classify sleep stages from a reduced set of wearable sensors, or entirely contact-free using a near-infrared camera.

**Part I** of this thesis developed deep neural networks to classify sleep stages first from heart rate and breathing rate time series in Chapter 3, then from underlying cardiac and respiratory waveforms in Chapter 4. Using a deep transfer learning approach, these models were successfully applied to physiological signals recorded entirely from a near-infrared video camera, enabling contact-free sleep monitoring. This was the first work to combine multiple physiological signals extracted from

video to perform automated sleep staging, and the importance of each signal to the overall performance of the model was demonstrated. The use of cardiac and respiratory waveforms resulted in an overall model-expert agreement that, at the time of writing, is significantly higher than any prior work that has addressed this problem. Meanwhile, the use of motion signals was shown to be important for the accurate classification of fragmented sleep, which is particularly important for expert interpretation of sleep hypnograms [221].

**Part II** of this thesis addressed two key issues with existing sleep staging methods. Firstly, although transfer learning was essential to the success of the vision-based methods introduced in **Part I**, it has several disadvantages which were highlighted in Chapter 5. To address these, a novel neural network architecture (wav2sleep) was introduced that could operate on an arbitrary set of input signals. After jointly training on an unprecedented amount of publicly available PSG data from multiple datasets with different available input signals, this single, unified model could subsequently be applied during testing to any subset of the input signals previously used during training. Doing so, it surpassed the sleep stage classification performance of existing methods across several input modalities of interest, such as the ECG, PPG and/or respiratory signals. The model has been open-sourced, enabling the wider research community to apply it to their data and to quickly build on these advances.

The improved performance obtained in Chapter 5 was enabled by the use of deep learning methods in combination with large quantities of data. However, Chapter 6 acknowledged that this trend may not continue and highlighted that consistent expert-level sleep staging performance from modalities such as the ECG or PPG, which can be recorded by modern wearable devices, may be an unachievable goal. By addressing flawed modelling assumptions used by most existing methods in the literature, and using techniques such as deep ensembling [117], this chapter introduced a model (SleepDecoder) that produced calibrated measures of sleep staging uncertainty. This was used to derive principled measures of uncertainty over specific points in time, over entire nights, and over derived sleep metrics. We

then demonstrated how these measures of uncertainty could be used in practice, for applications such as sleep disorder screening or evaluating treatment efficacy. Though Chapter 6 focused on sleep staging from the ECG, the methods introduced can be applied to any input modality.

## 7.2 Future Work

This section discusses future research directions that build on the work of this thesis. This covers potential improvements to sleep monitoring from wearables and video, wider applications of the models and methods introduced, and proposals for the clinical evaluation of these algorithms.

### 7.2.1 Vision-based sleep monitoring

**Improved motion representations.** In Chapter 3, we moved from a feature engineering to a deep learning approach using the vital signs (heart rate and breathing rate). Then, in Chapter 4, we moved from using heart and breathing rates as model inputs to instead using the underlying cardiac and respiratory waveforms. These changes were progressively shown to result in significant improvements in sleep staging accuracy. This mirrors the wider trend observed in numerous domains that, with sufficient data, deep neural networks can often learn more effective representations of input data for downstream tasks than human-designed features. Whilst this was successfully explored for cardio-respiratory signals, motion information was only utilised in Chapters 3 and 4 via a simple feature engineering approach. This is because there was no analogous motion signal available in the (abundant) contact sensor datasets with which to train a deep neural network. In the visual domain, recent work such as DINOv2 [154] has shown that *linear* classifiers applied to the outputs of self-supervised[1] models can achieve comparable performance to fully-supervised methods across a range of downstream tasks. This indicates that the low-dimensional representations produced by these models must encode abstract visual information. In our domain, this may include information

[1]Where the learning objective is derived from unlabelled data, e.g. next token prediction [171].

such as sleep pose. These models could be used as feature extractors, to extract semantic information directly from video input. This additional information could then be passed to a downstream classifier such as those used in Chapters 3 and 4, which could further improve sleep staging performance, or be used as part of a system for the automated detection of sleep movement disorders [251].

**Comprehensive sleep assessment.** The algorithms introduced in Chapters 3 and 4 demonstrated that a neural network trained to solve a task using contact sensor data could be transferred to similar physiological signals measured from video, with limited data. These techniques could be used for applications such as the detection of sleep apnoeas, to provide a more complete assessment of sleep quality. They could also be used to improve the performance of sleep staging algorithms from other contact-free modalities, such as radar or mattress-based sensors, which may be better suited for applications such as continuous, at-home monitoring.

### 7.2.2 Physiological representation learning

This thesis has focused on the supervised learning task of classifying sleep stages from a variety of input modalities, which was made possible by a vast quantity of expert-annotated data. However, a promising future direction of research is unsupervised learning of physiological representations that could be applied to downstream tasks with a much smaller quantity of labelled data, or without any labels at all.

**Better representations of signals.** In the audio (time-series) domain, the wav2vec [15, 187] series of models attained state-of-the-art speech recognition performance by first performing unsupervised pre-training using large unlabelled datasets, then fine-tuning on small amounts of labelled data. Given the costly annotation process of overnight PSG recordings, similar approaches could help to further improve the performance of the sleep staging algorithms developed in this thesis. Indeed, the wav2sleep model introduced in Chapter 5 would be particularly well-suited to popular unsupervised learning approaches such as cross-modal prediction [5, 127].

**Better representations of sleep.** Chapter 6 highlighted issues with the existing discretisation of (continuous) sleep into sleep stages. This human-defined clustering provides an interpretable unit of sleep, but a natural question is whether better clusters could be learnt automatically from the data. This is not a novel question, with research into alternative representations of the sleep-wake continuum dating back to the 1990s [159]. However, there have been two major advances in the research landscape since that original work. Firstly, with sufficient data, deep neural networks have repeatedly been shown to be capable of learning more effective representations from raw input data than shallow, heuristic-based methods or human-designed alternatives [67]. Secondly, the quantity of sleep data available to researchers has increased by several orders of magnitude thanks to resources such as the National Sleep Research Resource, used throughout this thesis. Together these could enable the discovery of representations (and clusters) of physiological states that are better predictors of the endpoints (biomarkers) we ultimately care about, similar to the (known) links between N3 sleep and dementia [121], or between REM sleep and depression [27].

Recent work has shown that 'foundation' models can learn generic representations directly from physiological signals such as ECG and PPG which can be used for downstream tasks such as the classification of cardiovascular conditions [2] or sleep-disordered breathing [215]. Meanwhile, self-supervised learning from EEG signals has recovered similar latent structures to the existing AASM sleep stages [20]. These approaches have so far only utilised short, 30-second segments of time series. Utilising longitudinal data, i.e. multi-day time series, could further improve the utility of such models, enabling them to learn representations of long-range physiological trends, such as circadian rhythms and sleep cycles.

### 7.2.3 Clinical validation

The algorithms developed in this thesis can automatically classify sleep stages with a level of model-expert agreement comparable to that observed between human experts. However, the performance of these algorithms still lags behind those

which use brain signals such as the EEG. Indeed, given the sources of uncertainty discussed in Chapter 6, it may not be possible to consistently obtain expert-level agreement across a clinical population from non-EEG sensors. However, this does not necessarily mean that these algorithms cannot have clinical utility. For example, despite having low specificity [139], wrist-based actigraphy is routinely used, and is effective, for broadly characterising longitudinal (multi-day) sleep-wake patterns [8].

Similarly, sleep staging from wearables or vision may already be sufficiently accurate to provide value in numerous clinical settings, without necessarily requiring performance on a par with expert-annotated PSG. For example, in the ICU, REM and deep sleep are both severely suppressed after major surgery [13]. If sleep staging from modalities such as wearables is sufficiently accurate to track REM cycles, it could aid the monitoring of patient recovery and inform clinical decisions such as when to discharge the patient. However, suitability for these clinical applications cannot be discovered solely from the analysis of retrospective, single-night studies as done in this thesis; it must be established in partnership with clinicians. A natural next step is, therefore, to embed these technologies into clinical studies, to establish their utility for applications not just in sleep medicine, but also for other areas of medicine where accurate sleep monitoring has until now not been possible due to the cumbersome nature of PSG.

## 7.3 Concluding Remarks

Through the development and application of new and existing deep learning techniques, this thesis has advanced the accuracy of automated sleep staging from a range of less obtrusive modalities than those required by traditional, manual sleep monitoring using polysomnography. By enabling cost-effective, easier-to-deploy, and longitudinal sleep monitoring, the algorithms introduced in this thesis can improve our ability to diagnose and treat sleep disorders and help to further our understanding of the link between sleep and physical and mental health.

# Appendices

# A

# Statements of Authorship

# Chapter 3

**Title of Paper:** Deep Learning-Enabled Sleep Staging From Vital Signs and Activity Measured Using a Near-Infrared Video Camera

**Publication Status:** Published

**Publication Details:** J. F. Carter, J. Jorge, B. Venugopal, O. Gibson, and L. Tarassenko, 'Deep Learning-Enabled Sleep Staging From Vital Signs and Activity Measured Using a Near-Infrared Video Camera', in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, Jun. 2023, pp. 5940–5949.

## Student Confirmation

**Student name:** Jonathan Carter

**Contribution to the paper:**

Devised the proposed methods. Developed the requisite software. Conducted all of the experiments. Wrote the initial manuscript.

**Signature:**

**Date:** 03/01/2025

## Supervisor Confirmation

By signing this Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

**Supervisor name and title:** Professor Lionel Tarassenko

**Supervisor comments:**

This was Jonathan's work from start to finish. My contribution was only to ask questions at various stages of the work, and then to review the draft paper - which required minimal edits.

**Signature:**

L. Tarassenko

**Date:**

# Chapter 4

**Title of Paper:** SleepVST: Sleep Staging from Near-Infrared Video Signals using Pre-Trained Transformers

**Publication Status:** Published.

**Publication Details:** J. F. Carter, J. Jorge, O. Gibson, and L. Tarassenko, 'SleepVST: Sleep Staging from Near-Infrared Video Signals using Pre-Trained Transformers', in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 12479–12489.

## Student Confirmation

**Student name:** Jonathan Carter

**Contribution to the paper:**

Devised the proposed methods. Developed the requisite software. Conducted all of the experiments. Wrote the initial manuscript.

**Signature:**

**Date:** 03/01/2025

## Supervisor Confirmation

By signing this Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

**Supervisor name and title:** Professor Lionel Tarassenko

**Supervisor comments:**

This was Jonathan's work from start to finish. My contribution was only to ask questions at various stages of the work, and then to review the draft paper - which required minimal edits.

**Signature:**

**Date:**

# Chapter 5

**Title of Paper:** wav2sleep: A Unified Multi-Modal Approach to Sleep Stage Classification from Physiological Signals

**Publication Status:** Accepted for Publication.

**Publication Details:** J. F. Carter and L. Tarassenko, 'wav2sleep: A Unified Multi-Modal Approach to Sleep Stage Classification from Physiological Signals', in Proceedings of the 4th Machine Learning for Health Symposium, PMLR, Dec. 2024

## Student Confirmation

**Student name:** Jonathan Carter

**Contribution to the paper:**

Devised the proposed methods. Developed the requisite software. Conducted all of the experiments. Wrote the initial manuscript.

**Signature:**

**Date:** 03/01/2025

## Supervisor Confirmation

By signing this Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

**Supervisor name and title:** Professor Lionel Tarassenko

**Supervisor comments:**

This was Jonathan's work from start to finish. My contribution was only to ask questions at various stages of the work, and then to review the draft paper - which required minimal edits.

**Signature:**

**Date:**

# B

# Polysomnography Databases

Across many fields, end-to-end deep learning approaches have been shown to outperform heuristic-based and classical machine learning approaches [67]. A key ingredient to this success has been a substantial increase in the quality and quantity of data available. To enable the application of deep learning methods to the problem of automated sleep staging, this thesis has extensively used polysomnography databases managed by the National Sleep Research Resource (NSRR) [260]. This appendix provides a brief description of each of these databases.

## B.1 Database descriptions

**Sleep Heart Health Study (SHHS, [170])**

This was a multi-centre cohort study that was originally intended to determine the cardiovascular and other consequences of sleep-disordered breathing. Initially, 6,441 participants undertook an unattended PSG exam, which usually took place at the participants' homes. 3,295 participants undertook a second PSG exam, meaning that the SHHS database contains nearly 10,000 PSG exams in total. SHHS is by far the largest database of PSG recordings used in this thesis and has been widely used to evaluate automated sleep staging algorithms in prior work e.g. [95, 163].

**Multi-Ethnic Study of Atherosclerosis (MESA, [44])**

This is a longitudinal investigation of risk factors associated with the development and progression of cardiovascular disease. As part of this wider investigation, 2,237 participants undertook an at-home PSG exam between 2010 and 2012. The study recruited black, white, Hispanic, and Chinese-American men and women who were aged between 45 and 84 at a baseline examination in 2000-2002.

**Wisconsin Sleep Cohort (WSC, [256])**

This is an ongoing longitudinal study of sleep disorders, with a particular focus on sleep apnea. The WSC study recorded a baseline in-laboratory PSG exam of 1,500 Wisconsin state employees aged between 30 and 60, with follow-up exams at four-year intervals.

**Cleveland Family Study (CFS, [177]),**

This study, which began in 1990, originally aimed to quantify the familial aggregation of sleep apnea. It recruited 2,284 individuals from 361 families who undertook an in-laboratory PSG exam on up to 4 occasions over 16 years. The last exam occurred in February 2006. At the time of writing, only data from the last exam is available from the NSRR, data from earlier visits is expected to be made available in the future.

**Cleveland Children's Sleep and Health Study (CCSHS, [180])**

This study recruited 907 children between 8 and 11 years old for an initial at-home PSG exam, with three longitudinal visits. The most recent visit took place between 2006 and 2010 and included an in-laboratory PSG exam. The cohort is a stratified random sample of full-term and preterm children.

**Childhood Adenotonsillectomy Trial (CHAT, [137])**

This was a randomized controlled trial which investigated whether children with mild to moderate obstructive sleep apnea who received early adenotonsillectomy showed greater levels of neuro-cognitive functioning versus the standard of care. In total, 1,447 children aged between 5 and 10 years old had initial in-laboratory

PSG exams for screening and 464 were then randomized to treatment with a follow-up PSG exam after 7 months.

**MrOS Sleep study [29]**

This study recruited men aged 65 and over who needed to be able to walk unassisted and must not have had a bilateral hip replacement. Between December 2003 and March 2005, 3,135 participants undertook a full PSG exam. The aim of the study was to understand the relationship between sleep architecture, sleep disordered breathing, and cognitive function in older men. In Chapters 5 and 6, MrOS is used as a completely held-out test set, as previously used by Jones *et al.* [97]. This means that no data from MrOS was used during model training or for hyper-parameter selection.

## B.2 Discussion

Collectively, these databases managed by the NSRR contain a wide range of variability across age, ethnicity, sex and physiological conditions. There is also variation in the available input signals, the make and model of the recording equipment, and in whether the PSG exam was performed at home or in a sleep laboratory. Using a combination of recordings from several of these databases, Jones *et al.* [97] constructed a test set that approximately matched the demographics from the 2022 US census, which is used in Chapters 5 and 6 to evaluate the algorithms introduced in this thesis.

One limitation of these databases is that they have all been collected within the US. This could introduce bias (demographics, equipment etc.) in the resulting algorithms which mean that they underperform when applied to subjects from other geographies. Alternative PSG databases have been collected, such as SIESTA [106] and SOMNIA [65] in Europe, and the Chinese Clinical Sleep Database [60] in China. However, the streamlined access procedure for the databases managed by the NSRR [260] currently makes them a much easier option for researchers.

# Bibliography


[1] How Sleep Works - How Much Sleep Is Enough? | NHLBI, NIH, 2022, https://www.nhlbi.nih.gov/health/sleep/how-much-sleep. 1

[2] Salar Abbaspourazad, Oussama Elachqar, Andrew C. Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale Training of Foundation Models for Wearable Biosignals, 2024, http://arxiv.org/abs/2312.05409. 4, 145

[3] H. W. Agnew Jr., Wilse B. Webb, and Robert L. Williams. The First Night Effect: An Eeg Studyof Sleep. *Psychophysiology*, 2(3):263–266, 1966. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.1966.tb02650.x. 13, 37

[4] Michael S. Aldrich and Barbara Jahnke. Diagnostic value of video-EEG polysomnography. *Neurology*, 41(7):1060–1060, 1991. Publisher: Wolters Kluwer. 32

[5] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-Supervised Learning by Cross-Modal Audio-Video Clustering. In *Advances in Neural Information Processing Systems*, pages 9758–9770. Curran Associates, Inc., 2020. 144

[6] American Academy of Sleep Medicine. The AASM Manual for the Scoring of Sleep and Associated Events - Summary of Updates in Version 2.6, 2020, https://aasm.org/wp-content/uploads/2020/01/Summary-of-Updates-in-v2.6-1.pdf. 10

[7] American Academy of Sleep Medicine. The AASM Manual for the Scoring of Sleep and Associated Events - Summary of Updates in Version 3, 2023, https://aasm.org/wp-content/uploads/2020/01/Summary-of-Updates-in-v2.6-1.pdf. 33

[8] Sonia Ancoli-Israel, Roger Cole, Cathy Alessi, Mark Chambers, William Moorcroft, and Charles P. Pollak. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*, 26(3):342–392, 2003. 146

[9] Thomas Andrillon, Yuval Nir, Richard J. Staba, Fabio Ferrarelli, Chiara Cirelli, Giulio Tononi, and Itzhak Fried. Sleep Spindles in Humans: Insights from Intracranial EEG and Unit Recordings. *The Journal of Neuroscience*, 31(49): 17821–17834, 2011. 11, 12

[10] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. pages 2425–2433, 2015. 28

[11] Apple Inc. Estimating Sleep Stages from Apple Watch. Technical report, 2023. 31

[12] Eugene Aserinsky and Nathaniel Kleitman. Regularly Occurring Periods of Eye Motility, and Concomitant Phenomena, During Sleep. *Science*, 118(3062):273–274, 1953. 7

[13] J Aurell and D Elmqvist. Sleep in the surgical intensive care unit: continuous polygraphic recording of sleep in nine patients receiving postoperative care. *British Medical Journal (Clinical research ed.)*, 290(6474):1029–1032, 1985. 146

[14] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization, 2016, `http://arxiv.org/abs/1607.06450`. 20, 106

[15] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, pages 12449–12460. Curran Associates, Inc., 2020. 18, 65, 97, 144

[16] Ying-Wen Bai, Wen-Tai Li, and Cheng-Hsiang Yeh. Design and implementation of an embedded monitor system for body breath detection by using image processing methods. In *2010 Digest of Technical Papers International Conference on Consumer Electronics (ICCE)*, pages 193–194, 2010. 26, 61

[17] Jessie P. Bakker, Marco Ross, Ray Vasko, Andreas Cerny, Pedro Fonseca, Jeff Jasko, Edmund Shaw, David P. White, and Peter Anderer. Estimating sleep stages using cardiorespiratory signals: validation of a novel algorithm across a wide range of sleep-disordered breathing severity. *Journal of Clinical Sleep Medicine*, 17(7): 1343–1354, 2021. 31, 63, 70, 88, 90, 104

[18] Jessie P Bakker, Marco Ross, Andreas Cerny, Ray Vasko, Edmund Shaw, Samuel Kuna, Ulysses J Magalang, Naresh M Punjabi, and Peter Anderer. Scoring sleep with artificial intelligence enables quantification of sleep stage ambiguity: hypnodensity based on multiple expert scorers and auto-scoring. *Sleep*, 46(2):zsac154, 2023. 115, 128

[19] Guha Balakrishnan, Fredo Durand, and John Guttag. Detecting Pulse from Head Motions in Video. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3430–3437, Portland, OR, USA, 2013. IEEE. 25, 61

[20] Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical EEG signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021. Publisher: IOP Publishing. 145

[21] Yujia Bao, Srinivasan Sivanandan, and Theofanis Karaletsos. Channel Vision Transformers: An Image Is Worth 1 x 16 x 16 Words. In *The Twelfth International Conference on Learning Representations*, 2024. 97, 99

[22] Robert J. Barry, Adam R. Clarke, Stuart J. Johnstone, and Jacqueline A. Rushby. Timing of caffeine's impact on autonomic and central nervous system measures: Clarification of arousal effects. *Biological Psychology*, 77(3):304–316, 2008. 8, 113

[23] Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xLSTM: Extended Long Short-Term Memory, 2024, `http://arxiv.org/abs/2405.04517`. 22

[24] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer, 2020, `http://arxiv.org/abs/2004.05150`. 20, 107, 120

[25] Selim R. Benbadis, Yinsheng Qu, Michael C. Perry, Dudley S. Dinner, and Hector Warnes. Interrater reliability of the multiple sleep latency test. *Electroencephalography and Clinical Neurophysiology*, 95(4):302–304, 1995. 13

[26] Richard B. Berry. Chapter 13 - Polysomnography, Portable Monitoring, and Actigraphy. In *Fundamentals of Sleep Medicine*, pages 189–218. W.B. Saunders, Saint Louis, 2012. 13, 28

[27] Richard B. Berry. Chapter 6 - Sleep Architecture Parameters, Normal Sleep, and Sleep Loss. In *Fundamentals of Sleep Medicine*, pages 79–90. W.B. Saunders, Saint Louis, 2012. 9, 145

[28] Christian Berthomier, Vincenzo Muto, Christina Schmidt, Gilles Vandewalle, Mathieu Jaspar, Jonathan Devillers, Giulia Gaggioni, Sarah L. Chellappa, Christelle Meyer, Christophe Phillips, Eric Salmon, Pierre Berthomier, Jacques Prado, Odile Benoit, Romain Bouet, Marie Brandewinder, Jérémie Mattout, and Pierre Maquet. Exploring scoring methods for research studies: Accuracy and variability of visual and automated sleep scoring. *Journal of Sleep Research*, 29(5):e12994, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jsr.12994. 13

[29] Terri Blackwell, Kristine Yaffe, Sonia Ancoli-Israel, Susan Redline, Kristine E. Ensrud, Marcia L. Stefanick, Alison Laffan, Katie L. Stone, and for the Osteoporotic Fractures in Men Study Group. Associations Between Sleep Architecture and Sleep-Disordered Breathing and Cognition in Older Community-Dwelling Men: The Osteoporotic Fractures in Men Sleep Study. *Journal of the American Geriatrics Society*, 59(12):2217–2225, 2011. 93, 155

[30] Guy E Blelloch. Prefix Sums and Their Applications. 22

[31] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, Pittsburgh Pennsylvania USA, 1992. ACM. 14

[32] Philippe Boudreau, Wei-Hsien Yeh, Guy A. Dumont, and Diane B. Boivin. Circadian Variation of Heart Rate Variability Across Sleep Stages. *Sleep*, 36(12):1919–1928, 2013. 8

[33] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. 24, 69

[34] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter,

Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, 2020, http://arxiv.org/abs/2005.14165. 22

[35] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995. Publisher: Society for Industrial and Applied Mathematics. 122

[36] Francesco P. Cappuccio, Lanfranco D'Elia, Pasquale Strazzullo, and Michelle A. Miller. Sleep Duration and All-Cause Mortality: A Systematic Review and Meta-Analysis of Prospective Studies. *Sleep*, 33(5):585–592, 2010. 1

[37] Jonathan F. Carter, João Jorge, Bindia Venugopal, Oliver Gibson, and Lionel Tarassenko. Deep Learning-Enabled Sleep Staging From Vital Signs and Activity Measured Using a Near-Infrared Video Camera. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5940–5949. IEEE, 2023. 10, 59, 62, 63, 69, 70, 72, 76, 83

[38] Jonathan F. Carter, João Jorge, Oliver Gibson, and Lionel Tarassenko. SleepVST: Sleep Staging from Near-Infrared Video Signals using Pre-Trained Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12489, 2024. 90, 104, 105

[39] CDC. FastStats: Sleep in Adults, 2024, https://www.cdc.gov/sleep/data-research/facts-stats/adults-sleep-facts-and-stats.html. 1

[40] Stanislas Chambon, Mathieu N. Galtier, Pierrick J. Arnal, Gilles Wainrib, and Alexandre Gramfort. A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(4):758–769, 2018. Conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering. 90

[41] Peter H. Charlton, Timothy Bonnici, Lionel Tarassenko, David A. Clifton, Richard Beale, and Peter J. Watkinson. An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram. *Physiological measurement*, 37(4):610, 2016. 30

[42] Peter H. Charlton, John Allen, Raquel Bailón, Stephanie Baker, Joachim A. Behar, Fei Chen, Gari D. Clifford, David A. Clifton, Harry J. Davies, and Cheng Ding. The 2023 wearable photoplethysmography roadmap. *Physiological measurement*, 44 (11):111001, 2023. 89

[43] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–365, 2018. 26, 61

[44] Xiaoli Chen, 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 Disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep*, 38(6):877–888, 2015. 13, 62, 70, 93, 154

[45] François Chollet. Xception: Deep Learning with Depthwise Separable Convolutions, 2017, `http://arxiv.org/abs/1610.02357`. 17

[46] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. 28, 38, 60, 69

[47] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996. 116

[48] Roy Cox, Winni F. Hofman, and Lucia M. Talamini. Involvement of spindles in memory consolidation is slow wave sleep-specific. *Learning & Memory*, 19(7): 264–267, 2012. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. 9

[49] Kris Cuppens, Lieven Lagae, Berten Ceulemans, Sabine Van Huffel, and Bart Vanrumste. Automatic video detection of body movement during sleep based on optical flow in pediatric patients with epilepsy. *Medical & Biological Engineering & Computing*, 48(9):923–931, 2010. 27, 40, 62

[50] Eliran Dafna, Ariel Tarasiuk, and Yaniv Zigel. Sleep staging using nocturnal sound analysis. *Scientific reports*, 8(1):1–14, 2018. Publisher: Nature Publishing Group. 32

[51] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 886–893 vol. 1, 2005. ISSN: 1063-6919. 14

[52] Heidi Danker-Hopfe, Peter Anderer, Josef Zeitlhofer, Marion Boeck, Hans Dorn, Georg Gruber, Esther Heller, Erna Loretz, Doris Moser, Silvia Parapatics, Bernd Saletu, Andrea Schmidt, and Georg Dorffner. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *Journal of Sleep Research*, 18(1):74–84, 2009. 28, 38, 61, 114

[53] Koldo De Miguel, Alberto Brunete, Miguel Hernando, and Ernesto Gambao. Home Camera-Based Fall Detection System for the Elderly. *Sensors*, 17(12):2864, 2017. 59

[54] Massimiliano de Zambotti, John Trinder, Alessandro Silvani, Ian M. Colrain, and Fiona C. Baker. Dynamic coupling between the central and autonomic nervous systems during sleep: A review. *Neuroscience & Biobehavioral Reviews*, 90:84–103, 2018. 113

[55] Lourdes M DelRosso, Caroline V Jackson, Kimberly Trotter, Oliviero Bruni, and Raffaele Ferri. Video-polysomnographic characterization of sleep movements in children with restless sleep disorder. *Sleep*, 42(4), 2019. 38

[56] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019, `http://arxiv.org/abs/1810.04805`. 91, 92

[57] Derk-Jan Dijk. Regulation and Functional Correlates of Slow Wave Sleep. *Journal of Clinical Sleep Medicine*, 5(2 Suppl):S6–S15, 2009. 11

[58] Feihong Ding, Andrew Cotton-Clay, Laura Fava, Venkat Easwar, Arthur Kinsolving, Philippe Kahn, Anil Rama, and Clete Kushida. Polysomnographic validation of an under-mattress monitoring device in estimating sleep architecture and obstructive sleep apnea in adults. *Sleep Medicine*, 96:20–27, 2022. 33

[59] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2020. 18, 22, 65, 92

[60] Ruichen Fang, Yihong Cheng, Fan Li, Yan Xu, Yuanhui Li, Xiang Liu, Simin Guo, Yuling Wang, Jinnong Jiang, Dan Zhou, and Bin Zhang. The Chinese Clinical Sleep Database: An Innovative Database System Includes Large-Scale Clinical Data of Chinese Population. *Nature and Science of Sleep*, Volume 16:305–313, 2024. 155

[61] Luigi Fiorillo, Paolo Favaro, and Francesca Dalia Faraci. DeepSleepNet-Lite: A Simplified Automatic Sleep Stage Scoring Model With Uncertainty Estimates. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:2076–2085, 2021. 118

[62] Luigi Fiorillo, Davide Pedroncelli, Valentina Agostini, Paolo Favaro, and Francesca Dalia Faraci. Multi-scored sleep databases: how to exploit the multiple-labels in automated sleep scoring. *Sleep*, 46(5):zsad028, 2023. 114

[63] T. B. Fitzpatrick. The validity and practicality of sun-reactive skin types I through VI. *Archives of Dermatology*, 124(6):869–871, 1988. 41, 72

[64] Pedro Fonseca, Xi Long, Mustafa Radha, Reinder Haakma, Ronald M. Aarts, and Jérôme Rolink. Sleep stage classification with ECG and respiratory effort. *Physiological Measurement*, 36(10):2027–2040, 2015. 31

[65] Merel M. van Gilst, Johannes P. van Dijk, Roy Krijn, Bertram Hoondert, Pedro Fonseca, Ruud J. G. van Sloun, Bruno Arsenali, Nele Vandenbussche, Sigrid Pillen, Henning Maass, Leonie van den Heuvel, Reinder Haakma, Tim R. Leufkens, Coen Lauwerijssen, Jan W. M. Bergmans, Dirk Pevernagie, and Sebastiaan Overeem. Protocol of the SOMNIA project: an observational study to create a neurophysiological database for advanced clinical sleep monitoring. *BMJ Open*, 9 (11):e030996, 2019. Publisher: British Medical Journal Publishing Group Section: Research methods. 155

[66] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One Embedding Space To Bind Them All, 2023, `http://arxiv.org/abs/2305.05665`. 22

[67] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. 4, 14, 90, 145, 153

[68] Trisha Greenhalgh, Glenn Robert, Fraser Macfarlane, Paul Bate, and Olivia Kyriakidou. Diffusion of Innovations in Service Organizations: Systematic Review and Recommendations. *The Milbank Quarterly*, 82(4):581–629, 2004. 115

[69] Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces, 2023, `http://arxiv.org/abs/2312.00752`. 22

[70] Alessandro R. Guazzi, Mauricio Villarroel, João Jorge, Jonathan Daly, Matthew C. Frise, Peter A. Robbins, and Lionel Tarassenko. Non-contact measurement of oxygen saturation with an RGB camera. *Biomedical Optics Express*, 6(9):3320–3338, 2015. Publisher: Optica Publishing Group. 27

[71] Antoine Guillot, Fabien Sauvet, Emmanuel H. During, and Valentin Thorey. Dreem Open Datasets: Multi-Scored Sleep Datasets to Compare Human and Automated Sleep Staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(9):1955–1965, 2020. Conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering. 111, 114, 124

[72] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 111, 118, 121

[73] Guillermo Gutierrez, Jeffrey Williams, Ghadah A. Alrehaili, Anna McLean, Ramin Pirouz, Richard Amdur, Vivek Jain, Jalil Ahari, Amandeep Bawa, and Shawn Kimbro. Respiratory rate variability in sleeping adults without obstructive sleep apnea. *Physiological Reports*, 4(17):e12949, 2016. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.14814/phy2.12949. 8

[74] Marco Hafner, Martin Stepanek, Jirka Taylor, Wendy M. Troxel, and Christian van Stolk. Why Sleep Matters—The Economic Costs of Insufficient Sleep. *Rand Health Quarterly*, 6(4):11, 2017. 58

[75] Péter Halász, Mario Terzano, Liborio Parrino, and Róbert Bódizs. The nature of arousal in sleep. *Journal of Sleep Research*, 13(1):1–23, 2004. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2869.2004.00388.x. 8

[76] M. Harford, J. Catherall, S. Gerry, J. D. Young, and P. Watkinson. Availability and performance of image-based, non-contact methods of monitoring heart rate, blood pressure, respiratory rate, and oxygen saturation: a systematic review. *Physiological Measurement*, 40(6):06TR01, 2019. Publisher: IOP Publishing. 28

[77] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 39, 42, 67

[78] Kaiming He, 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*, pages 770–778, 2016. 18, 20, 47, 65, 92

[79] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners, 2021, `http://arxiv.org/abs/2111.06377`. 97

[80] Adrienne Heinrich, Xavier Aubert, and Gerard de Haan. Body movement analysis during sleep based on video motion estimation. In *2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013)*, pages 539–543, 2013. 40, 59

[81] Adrienne Heinrich, Di Geng, Dmitry Znamenskiy, Jelte Peter Vink, and Gerard de Haan. Robust and Sensitive Video Motion Detection for Sleep Analysis. *IEEE Journal of Biomedical and Health Informatics*, 18(3):790–798, 2014. Conference Name: IEEE Journal of Biomedical and Health Informatics. 40

[82] Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs), 2023, `http://arxiv.org/abs/1606.08415`. 96

[83] Ellen Herbst, Thomas J. Metzler, Maryann Lenoci, Shannon E. McCaslin, Sabra Inslicht, Charles R. Marmar, and Thomas C. Neylan. Adaptation Effects to Sleep Studies in Participants with and without Chronic Posttraumatic Stress Disorder. *Psychophysiology*, 47(6):1127–1133, 2010. 13

[84] Elisabeth Heremans, Laura Van den Bulcke, Nabeel Seedat, Astrid Devulder, Pascal Borzée, Bertien Buyse, Dries Testelmans, Maarten Van Den Bossche, Mihaela van der Schaar, and Maarten De Vos. Automated remote sleep monitoring needs machine learning with uncertainty quantification. 2023. 118

[85] Elisabeth R. M. Heremans, Nabeel Seedat, Bertien Buyse, Dries Testelmans, Mihaela van der Schaar, and Maarten De Vos. U-PASS: an Uncertainty-guided deep learning Pipeline for Automated Sleep Staging, 2023, `http://arxiv.org/abs/2306.04663`. 117

[86] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. Publisher: MIT press. 22, 31, 119

[87] Hong Hong, Li Zhang, Chen Gu, Yusheng Li, Guangxin Zhou, and Xiaohua Zhu. Noncontact sleep stage estimation using a CW Doppler radar. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 8(2):260–270, 2018. Publisher: IEEE. 32

[88] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1):185–203, 1981. 39, 42

[89] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. 15

[90] Yukai Huang, Dongmin Huang, Jia Huang, Hongzhou Lu, Min He, and Wenjin Wang. Camera Wavelength Selection for Multi-wavelength Pulse Transit Time based Blood Pressure Monitoring. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2023:1–5, 2023. 27

[91] D. W. Hudgel, R. J. Martin, B. Johnson, and P. Hill. Mechanics of the respiratory system and breathing pattern during sleep in normal humans. *Journal of Applied Physiology: Respiratory, Environmental and Exercise Physiology*, 56(1):133–137, 1984. 30, 38, 59, 62

[92] Nicholas Dunkley Hutchinson and Simon Mark Chave Jones. Method and apparatus for monitoring of a human or animal subject field. US11563920B2, 2023. 44

[93] Isabel C. Hutchison and Shailendra Rathore. The role of REM sleep theta activity in emotional memory. *Frontiers in Psychology*, 6, 2015. Publisher: Frontiers. 9

[94] C. Iber. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology, and Technical Specification. 2007. 8, 59, 61, 86, 113

[95] Syed Anas Imtiaz. A systematic review of sensing technologies for wearable sleep staging. *Sensors*, 21(5):1562, 2021. Publisher: MDPI. 153

[96] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 81, 106

[97] Adam M. Jones, Laurent Itti, and Bhavin R. Sheth. Expert-level sleep staging using an electrocardiography-only feed-forward neural network. *Computers in Biology and Medicine*, 176:108545, 2024. 94, 95, 103, 104, 106, 124, 136, 155

[98] João Jorge, Mauricio Villarroel, Hamish Tomlinson, Oliver Gibson, Julie L. Darbyshire, Jody Ede, Mirae Harford, John Duncan Young, Lionel Tarassenko, and Peter Watkinson. Non-contact physiological monitoring of post-operative patients in the intensive care unit. *npj Digital Medicine*, 5(1):4, 2022. 26, 31, 44

[99] M. Jouvet. Paradoxical Sleep — A Study of its Nature and Mechanisms*. In *Progress in Brain Research*, pages 20–62. Elsevier, 1965. 7

[100] Masamitsu Kamon, Shima Okada, Masafumi Furuta, and Koki Yoshida. Development of a non-contact sleep monitoring system for children. *Frontiers in Digital Health*, 4, 2022. 4, 32, 62, 70

[101] Dae Y. Kang, Pamela N. DeYoung, Justin Tantiongloc, Todd P. Coleman, and Robert L. Owens. Statistical uncertainty quantification to augment clinical decision support: a first implementation in sleep medicine. *npj Digital Medicine*, 4(1):1–9, 2021. Number: 1 Publisher: Nature Publishing Group. 117

[102] Jan W. Kantelhardt, Thomas Penzel, Sven Rostig, Heinrich F. Becker, Shlomo Havlin, and Armin Bunde. Breathing during REM and non-REM sleep: correlated versus uncorrelated behaviour. *Physica A: Statistical Mechanics and its Applications*, 319:447–457, 2003. 89

[103] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. Measuring Catastrophic Forgetting in Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. Number: 1. 90

[104] Sirvan Khalighi, Teresa Sousa, Gabriel Pires, and Urbano Nunes. Automatic sleep staging: A computer assisted approach for optimal combination of features and polysomnographic channels. *Expert Systems with Applications*, 40(17):7046–7059, 2013. 12

[105] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 16, 54

[106] G. Klosh, B. Kemp, T. Penzel, A. Schlogl, P. Rappelsberger, E. Trenker, G. Gruber, J. Zeithofer, B. Saletu, W.M. Herrmann, S.L. Himanen, D. Kunz, M.J. Barbanoj, J. Roschke, A. Varri, and G. Dorffner. The SIESTA project polygraphic and clinical database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):51–57, 2001. Conference Name: IEEE Engineering in Medicine and Biology Magazine. 31, 155

[107] Kohlschütter, E. Messungen der Festigkeit des Schlafes. *Z Ration Med.*, 1863. 7

[108] Konstantinos Kontras, Christos Chatzichristos, Huy Phan, Johan Suykens, and Maarten De Vos. CoRe-Sleep: A Multimodal Fusion Framework for Time Series Robust to Imperfect Modalities. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:840–849, 2024. Conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering. 90

[109] Henri Korkalainen, Juhani Aakko, Sami Nikkonen, Samu Kainulainen, Akseli Leino, Brett Duce, Isaac O. Afara, Sami Myllymaa, Juha Töyräs, and Timo Leppänen. Accurate Deep Learning-Based Sleep Staging in a Clinical Population With Suspected Obstructive Sleep Apnea. *IEEE Journal of Biomedical and Health Informatics*, 24(7):2073–2081, 2020. 101

[110] Kevin Kotzen, Peter H. Charlton, Sharon Salabi, Lea Amar, Amir Landesberg, and Joachim A. Behar. SleepPPG-Net: A Deep Learning Algorithm for Robust Sleep Staging From Continuous Photoplethysmography. *IEEE Journal of Biomedical and Health Informatics*, 27(2):924–932, 2023. 28, 31, 59, 61, 69, 70, 88, 89, 92, 95, 98, 107, 119, 130, 132

[111] Till Kroeger, Radu Timofte, Dengxin Dai, and Luc Van Gool. Fast Optical Flow using Dense Inverse Search, 2016. 42, 83

[112] James M. Krueger, Joseph T. Nguyen, Cheryl J. Dykstra-Aiello, and Ping Taishi. Local sleep. *Sleep Medicine Reviews*, 43:14–21, 2019. 33

[113] Bjorn Kruger, Anna Vogele, Marouane Lassiri, Lukas Herwartz, Thomas Terkatz, and Andreas Weber. Sleep detection using de-identified depth data. *Journal of Mobile Multimedia*, pages 327–342, 2014. 40

[114] Andrew D. Krystal. Psychiatric disorders and sleep. *Neurologic clinics*, 30(4): 1389–1413, 2012. Publisher: Elsevier. 113

[115] Clete A. Kushida, Michael R. Littner, Timothy Morgenthaler, Cathy A. Alessi, Dennis Bailey, Jack Coleman, Leah Friedman, Max Hirshkowitz, Sheldon Kapen, Milton Kramer, Teofilo Lee-Chiong, Daniel L. Loube, Judith Owens, Jeffrey P. Pancer, and Merrill Wise. Practice parameters for the indications for polysomnography and related procedures: an update for 2005. *Sleep*, 28(4):499–521, 2005. 9

[116] Hyun Bin Kwon, Sang Ho Choi, Dongseok Lee, Dongyeon Son, Heenam Yoon, Mi Hyun Lee, Yu Jin Lee, and Kwang Suk Park. Attention-Based LSTM for Non-Contact Sleep Stage Classification Using IR-UWB Radar. *IEEE Journal of Biomedical and Health Informatics*, 25(10):3844–3853, 2021. Conference Name: IEEE Journal of Biomedical and Health Informatics. 32

[117] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 111, 120, 142

[118] J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977. Publisher: [Wiley, International Biometric Society]. 29, 100

[119] Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995. Publisher: Citeseer. 16

[120] Ellen E. Lee, Sonia Ancoli-Israel, Lisa T. Eyler, Xin M. Tu, Barton W. Palmer, Michael R. Irwin, and Dilip V. Jeste. Sleep Disturbances and Inflammatory Biomarkers in Schizophrenia: Focus on Sex Differences. *The American Journal of Geriatric Psychiatry: Official Journal of the American Association for Geriatric Psychiatry*, 27(1):21–31, 2019. 2

[121] Yee Fun Lee, Dmitry Gerashchenko, Igor Timofeev, Brian J. Bacskai, and Ksenia V. Kastanenka. Slow Wave Sleep Is a Promising Intervention Target for Alzheimer's Disease. *Frontiers in Neuroscience*, 14, 2020. 9, 58, 145

[122] Yun Ji Lee, Jae Yong Lee, Jae Hoon Cho, and Ji Ho Choi. Interrater reliability of sleep stage scoring: a meta-analysis. *Journal of Clinical Sleep Medicine*, 18(1): 193–202, 2022. 61

[123] Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. xFormers: A modular and hackable Transformer modelling library, 2022, `https://github.com/facebookresearch/xformers`. 107

[124] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jędrzej Nowak. Measuring pulse rate with a webcam — A non-contact method for evaluating cardiac activity. In *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 405–410, 2011. 44

[125] Michael H. Li, Azadeh Yadollahi, and Babak Taati. A non-contact vision-based system for respiratory rate estimation. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2119–2122, 2014. 26, 44, 61

[126] Q. Li, R. G. Mark, and G. D. Clifford. Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter. *Physiological Measurement*, 29(1):15–32, 2007. Publisher: IOP Publishing. 44

[127] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10609–10618, 2019. 144

[128] Jia-Wei Lin, Ming-Hung Lu, and Yuan-Hsiang Lin. A Thermal Camera Based Continuous Body Temperature Measurement System. pages 0–0, 2019. 27

[129] Xin Liu, Brian Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. EfficientPhys: Enabling Simple, Fast and Accurate Camera-Based Cardiac Measurement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5008–5017, 2023. 61

[130] Hugh Lloyd-Jukes, Oliver John Gibson, Tracey Wrench, Ade Odunlade, and Lionel Tarassenko. Vision-Based Patient Monitoring and Management in Mental Health Settings. *Journal of Clinical Engineering*, 46(1):36–43, 2021. 27

[131] Xi Long, Renée Otte, Eric van der Sanden, Jan Werth, and Tao Tan. Video-Based Actigraphy for Monitoring Wake and Sleep in Healthy Infants: A Laboratory Study. *Sensors*, 19(5):1075, 2019. 27, 38, 40

[132] Xi Long, Javier Espina, Renée A. Otte, Wenjin Wang, Ronald M. Aarts, and Peter Andriessen. Video-based actigraphy is an effective contact-free method of assessing sleep in preterm infants. *Acta Paediatrica (Oslo, Norway : 1992)*, 110(6):1815–1816, 2021. 27

[133] Dominique Lorrain and Joseph De Koninck. Sleep Position and Sleep Stages: Evidence of Their Independence. *Sleep*, 21(4):335–340, 1998. 28

[134] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. 66, 95, 124

[135] Bruce D Lucas and Takeo Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proceedings of the 7th Joint International Conference on Artificial Intelligence (IJCAI)*, page 9, 1981. 25

[136] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a Closer Look at Domain Shift: Category-Level Adversaries for Semantics Consistent Domain Adaptation. pages 2507–2516, 2019. 62

[137] Carole L. Marcus, Reneé H. Moore, Carol L. Rosen, Bruno Giordani, Susan L. Garetz, H. Gerry Taylor, Ron B. Mitchell, Raouf Amin, Eliot S. Katz, Raanan Arens, Shalini Paruthi, Hiren Muzumdar, David Gozal, Nina Hattiangadi Thomas, Janice Ware, Dean Beebe, Karen Snyder, Lisa Elden, Robert C. Sprecher, Paul Willging, Dwight Jones, John P. Bent, Timothy Hoban, Ronald D. Chervin, Susan S. Ellenberg, Susan Redline, and Childhood Adenotonsillectomy Trial (CHAT). A randomized trial of adenotonsillectomy for childhood sleep apnea. *The New England Journal of Medicine*, 368(25):2366–2376, 2013. 93, 154

[138] ELWIN MARG. DEVELOPMENT OF ELECTRO-OCULOGRAPHY: Standing Potential of the Eye in Registration of Eye Movement. *A.M.A. Archives of Ophthalmology*, 45(2):169–185, 1951. 7

[139] Miguel Marino, Yi Li, Michael N. Rueschman, J. W. Winkelman, J. M. Ellenbogen, J. M. Solet, Hilary Dulin, Lisa F. Berkman, and Orfeu M. Buxton. Measuring Sleep: Accuracy, Sensitivity, and Specificity of Wrist Actigraphy Compared to Polysomnography. *Sleep*, 36(11):1747–1755, 2013. 14, 146

[140] S E Martin, H M Engleman, I J Deary, and N J Douglas. The effect of sleep fragmentation on daytime function. *American Journal of Respiratory and Critical Care Medicine*, 153(4):1328–1332, 1996. Publisher: American Thoracic Society - AJRCCM. 73

[141] Daniel McDuff. Camera Measurement of Physiological Vital Signs. *ACM Computing Surveys*, 55(9):1–40, 2023. 24, 25, 31, 61

[142] Institute of Medicine, Board on Health Sciences Policy, and Committee on Sleep Medicine and Research. *Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem*. National Academies Press, 2006. Google-Books-ID: 3bVTAgAAQBAJ. 1

[143] Sara Mahvash Mohammadi, Majdi Alnowami, Sofia Khan, Derk-Jan Dijk, Adrian Hilton, and Kevin Wells. Sleep Posture Classification using a Convolutional Neural Network. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4, 2018. ISSN: 1558-4615. 28

[144] Sara Mahvash Mohammadi, Shirin Enshaeifar, Adrian Hilton, Derk-Jan Dijk, and Kevin Wells. Transfer Learning for Clinical Sleep Pose Detection Using a Single 2D IR Camera. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:290–299, 2021. 42, 53, 67

[145] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 15

[146] K. Nakajima, A. Osa, and H. Miike. A method for measuring respiration and physical activity in bed by optical flow analysis. In *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.*, pages 2054–2057 vol.5, 1997. 26, 61

[147] Kazuki Nakajima, Yoshiaki Matsumoto, and Toshiyo Tamura. Development of real-time image sequence analysis for evaluating posture change and respiratory rate of a subject in bed. *Physiological Measurement*, 22(3):N21–N28, 2001. 26, 27, 40, 44, 61, 62

[148] Borum Nam, Beomjun Bark, Jeyeon Lee, and In Young Kim. InsightSleepNet: the interpretable and uncertainty-aware deep learning network for sleep staging using continuous Photoplethysmography. *BMC Medical Informatics and Decision Making*, 24(1):50, 2024. 117, 118

[149] Chetan S. Nayak and Arayamparambil C. Anilkumar. *EEG Normal Sleep*. StatPearls Publishing, 2022. Publication Title: StatPearls [Internet]. 11, 12

[150] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*, pages 625–632, Bonn, Germany, 2005. ACM Press. 126

[151] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *The Eleventh International Conference on Learning Representations*, 2022. 64, 65

[152] E. Nippolainen, N. P. Podolian, R. V. Romashko, Y. N. Kulchin, and A. A. Kamshilin. Photoplethysmographic Waveform as a Function of Subject's Age. *Physics Procedia*, 73:241–245, 2015. 30

[153] Teruaki Nochino, Yuko Ohno, Takafumi Kato, Masako Taniike, and Shima Okada. Sleep stage estimation method using a camera for home use. *Biomedical Engineering Letters*, 9(2):257–265, 2019. 4, 32, 40, 43, 49, 59, 62, 66, 70

[154] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, 2024, `http://arxiv.org/abs/2304.07193`. 143

[155] Christina Orphanidou, Timothy Bonnici, Peter Charlton, David Clifton, David Vallance, and Lionel Tarassenko. Signal-Quality Indices for the Electrocardiogram and Photoplethysmogram: Derivation and Applications to Wireless Monitoring. *IEEE Journal of Biomedical and Health Informatics*, 19(3):832–838, 2015. 112

[156] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model' s uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 118, 120

[157] J. F. Pagel and Bennett L. Parnes. Medications for the treatment of sleep disorders: an overview. *Primary care companion to the Journal of clinical psychiatry*, 3(3): 118, 2001. Publisher: Physicians Postgraduate Press, Inc. 8, 113

[158] Jiapu Pan and Willis J. Tompkins. A Real-Time QRS Detection Algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3):230–236, 1985. 53, 62

[159] James Pardey, Stephen Roberts, Lionel Tarassenko, and John Stradling. A new approach to the analysis of the human sleep/wakefulness continuum. *Journal of sleep research*, 5(4):201–210, 1996. 59, 145

[160] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, 2019. 22, 81, 107

[161] Aakash K. Patel, Vamsi Reddy, and John F. Araujo. *Physiology, Sleep Stages*. StatPearls Publishing, 2022. 8, 66, 92

[162] T. Penzel. Is heart rate variability the simple solution to diagnose sleep apnoea? *European Respiratory Journal*, 22(6):870–971, 2003. Publisher: European Respiratory Society Section: Editorials. 101

[163] Huy Phan and Kaare Mikkelsen. Automatic sleep staging of EEG signals: recent development, challenges, and future directions. *Physiological Measurement*, 43(4): 04TR01, 2022. 13, 30, 33, 38, 45, 60, 88, 153

[164] Huy Phan, 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. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):400–410, 2019. 90

[165] Huy Phan, Oliver Y. Chén, Minh C. Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. XSleepNet: Multi-View Sequential Model for Automatic Sleep Staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 5903–5915, 2022. 59, 94, 114

[166] Huy Phan, Kaare Mikkelsen, Oliver Y. Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. SleepTransformer: Automatic Sleep Staging With Interpretability and Uncertainty Quantification. *IEEE Transactions on Biomedical Engineering*, 69 (8):2456–2467, 2022. 30, 38, 52, 66, 117, 118

[167] Nicolò Pini, Ju Lynn Ong, Gizem Yilmaz, Nicholas I. Y. N. Chee, Zhao Siting, Animesh Awasthi, Siddharth Biju, Kishan Kishan, Amiya Patanaik, William P. Fifer, and Maristella Lucchini. An automated heart rate-based algorithm for sleep stage classification: Validation using conventional polysomnography and an innovative wearable electrocardiogram device. *Frontiers in Neuroscience*, 16:974192, 2022. 38

[168] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam. *IEEE Transactions on Biomedical Engineering*, 58(1):7–11, 2011. Conference Name: IEEE Transactions on Biomedical Engineering. 25, 44, 61

[169] Jathurshan Pradeepkumar, Mithunjha Anandakumar, Vinith Kugathasan, Dhinesh Suntharalingham, Simon L. Kappel, Anjula C. De Silva, and Chamira U. S. Edussooriya. Towards Interpretable Sleep Stage Classification Using Cross-Modal Transformers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pages 1–1, 2024. 90

[170] S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet, and P. W. Wahl. The Sleep Heart Health Study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997. 13, 38, 45, 46, 62, 70, 93, 153

[171] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. Publisher: OpenAI. 143

[172] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1 (8):9, 2019. 18

[173] Mustafa Radha, Pedro Fonseca, Arnaud Moreau, Marco Ross, Andreas Cerny, Peter Anderer, Xi Long, and Ronald M. Aarts. Sleep stage classification from heart-rate variability using long short-term memory neural networks. *Scientific Reports*, 9(1): 14149, 2019. 31, 49

[174] Mustafa Radha, Pedro Fonseca, Arnaud Moreau, Marco Ross, Andreas Cerny, Peter Anderer, Xi Long, and Ronald M. Aarts. A deep transfer learning approach for wearable sleep stage classification with photoplethysmography. *npj Digital Medicine*, 4(1):1–11, 2021. 28, 31, 34, 38, 45, 61, 89

[175] Rahul Rahaman and Alexandre H. Thiery. Uncertainty quantification and deep ensembles. *Advances in Neural Information Processing Systems*, 34:20063–20075, 2021. 121

[176] Rechtschaffen, A. and Kales, A. A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects., 1968. 8

[177] S. Redline, P. V. Tishler, T. D. Tosteson, J. Williamson, K. Kump, I. Browner, V. Ferrette, and P. Krejci. The familial aggregation of obstructive sleep apnea. *American Journal of Respiratory and Critical Care Medicine*, 151(3 Pt 1):682–687, 1995. 93, 154

[178] Susan Redline, Gayane Yenokyan, Daniel J. Gottlieb, Eyal Shahar, George T. O'Connor, Helaine E. Resnick, Marie Diener-West, Mark H. Sanders, Philip A. Wolf, Estella M. Geraghty, Tauqeer Ali, Michael Lebowitz, and Naresh M. Punjabi. Obstructive Sleep Apnea–Hypopnea and Incident Stroke. *American Journal of Respiratory and Critical Care Medicine*, 182(2):269–277, 2010. Publisher: American Thoracic Society - AJRCCM. 13

[179] S.J. Redmond and C. Heneghan. Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea. *IEEE Transactions on Biomedical Engineering*, 53(3): 485–496, 2006. 30, 31, 59

[180] Carol L. Rosen, 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. *The Journal of Pediatrics*, 142(4): 383–389, 2003. 93, 154

[181] Richard S. Rosenberg and Hout Steven Van. The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring. *Journal of Clinical Sleep Medicine*, 09(01):81–87. Publisher: American Academy of Sleep Medicine. 13

[182] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958. Publisher: American Psychological Association. 14

[183] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986. Publisher: Nature Publishing Group UK London. 14, 16

[184] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, 2015, `http://arxiv.org/abs/1409.0575`. 14

[185] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution Strategies as a Scalable Alternative to Reinforcement Learning, 2017, `http://arxiv.org/abs/1703.03864`. 16

[186] Fons Schipper, Angela Grassi, Marco Ross, Andreas Cerny, Peter Anderer, Lieke Hermans, Fokke van Meulen, Mickey Leentjens, Emily Schoustra, Pien Bosschieter, Ruud J. G. van Sloun, Sebastiaan Overeem, and Pedro Fonseca. Overnight Sleep Staging Using Chest-Worn Accelerometry. *Sensors*, 24(17):5717, 2024. Number: 17 Publisher: Multidisciplinary Digital Publishing Institute. 31

[187] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised Pre-training for Speech Recognition, 2019, `http://arxiv.org/abs/1904.05862`. 144

[188] Hartmut Schulz. Rethinking Sleep Analysis. *Journal of Clinical Sleep Medicine*, 04 (02):99–103, 2008. Publisher: American Academy of Sleep Medicine. 33

[189] Hartmut Schulz and Piero Salzarulo. The Development of Sleep Medicine: A Historical Sketch. *Journal of Clinical Sleep Medicine*, 12(07):1041–1052, 2016. 7

[190] William J. Schwartz and Elizabeth B. Klerman. Circadian Neurobiology and the Physiological Regulation of Sleep and Wakefulness. *Neurologic clinics*, 37(3):475–486, 2019. 1

[191] A. J. Schwichtenberg, Jeehyun Choe, Ashleigh Kellerman, Emily A. Abel, and Edward J. Delp. Pediatric Videosomnography: Can Signal/Video Processing Distinguish Sleep and Wake States? *Frontiers in Pediatrics*, 6:158, 2018. 27, 40, 62

[192] Michael K. Scullin and Chenlu Gao. Dynamic Contributions of Slow Wave Sleep and REM Sleep to Cognitive Longevity. *Current Sleep Medicine Reports*, 4(4): 284–293, 2018. 13

[193] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014. 23

[194] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. 62

[195] Z. Shinar, A. Baharav, Y. Dagan, and S. Akselrod. Automatic detection of slow-wave-sleep using heart rate variability. In *Computers in Cardiology 2001. Vol.28*, pages 593–596, 2001. 30, 37, 38, 59, 62, 89

[196] Francesca Siclari and Giulio Tononi. Local aspects of sleep and wakefulness. *Current Opinion in Neurobiology*, 44:222–227, 2017. 33

[197] Alessandro Silvani, Raffaele Ferri, Viviana Lo Martire, Stefano Bastianini, Chiara Berteotti, Agnese Salvadè, Giuseppe Plazzi, Marco Zucconi, Luigi Ferini-Strambi, Claudio L. Bassetti, Mauro Manconi, and Giovanna Zoccoli. Muscle Activity During Sleep in Human Subjects, Rats, and Mice: Towards Translational Models of REM Sleep Without Atonia. *Sleep*, 40(4):zsx029, 2017. 8

[198] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 18

[199] Péter Simor, Gwen van der Wijk, Lino Nobili, and Philippe Peigneux. The microstructure of REM sleep: Why phasic and tonic? *Sleep Medicine Reviews*, 52: 101305, 2020. 33

[200] Farzad Siyahjani, Gary Garcia Molina, Shawn Barr, and Faisal Mushtaq. Performance Evaluation of a Smart Bed Technology against Polysomnography. *Sensors*, 22 (7):2605, 2022. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute. 33

[201] Freddie Bickford Smith, Jannik Kossen, Eleanor Trollope, Mark van der Wilk, Adam Foster, and Tom Rainforth. Rethinking Aleatoric and Epistemic Uncertainty. 2024. 112

[202] Jeremy Speth, Nathan Vance, Patrick Flynn, and Adam Czajka. Non-contrastive unsupervised learning of physiological signals from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14464–14474, 2023. 26, 61

[203] Niranjan Sridhar, Ali Shoeb, Philip Stephens, Alaa Kharbouch, David Ben Shimol, Joshua Burkart, Atiyeh Ghoreyshi, and Lance Myers. Deep learning for automated sleep staging using instantaneous heart rate. *npj Digital Medicine*, 3(1):1–10, 2020. 31, 61, 70, 88, 92, 94, 104, 107

[204] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. 20

[205] Neil Stanley. The Future of Sleep Staging, Revisited. *Nature and Science of Sleep*, 15:313–322, 2023. 33

[206] Ambra Stefani and Birgit Högl. Sleep in Parkinson's disease. *Neuropsychopharmacology*, 45(1):121–128, 2020. 2, 32, 38, 59, 75, 76

[207] Ambra Stefani, David Gabelia, Thomas Mitterling, Werner Poewe, Birgit Högl, and Birgit Frauscher. A Prospective Video-Polysomnographic Analysis of Movements during Physiological Sleep in 100 Healthy Sleepers. *Sleep*, 38(9):1479–1487, 2015. 10, 37, 58

[208] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024. 107, 119

[209] Haoqi Sun, Wolfgang Ganglberger, Ezhil Panneerselvam, Michael J Leone, Syed A Quadri, Balaji Goparaju, Ryan A Tesh, Oluwaseun Akeju, Robert J Thomas, and M Brandon Westover. Sleep staging from electrocardiography and respiration with deep learning. *Sleep*, 43(7):zsz306, 2020. 28, 31, 72

[210] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017. 114

[211] Hiroyuki Suzuki, Makoto Uchiyama, Hirokuni Tagaya, Akiko Ozaki, Kenichi Kuriyama, Sayaka Aritake, Kayo Shibui, Xin Tan, Yuichi Kamei, and Ryuichi Kuga. Dreaming During Non-rapid Eye Movement Sleep in the Absence of Prior Rapid Eye Movement Sleep. *Sleep*, 27(8):1486–1490, 2004. 7

[212] K. S. Tan, R. Saatchi, H. Elphick, and D. Burke. Real-time vision based respiration monitoring system. In *2010 7th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP 2010)*, pages 770–774, 2010. 26, 61

[213] Chameleon Team. Chameleon: Mixed-Modal Early-Fusion Foundation Models, 2024, `http://arxiv.org/abs/2405.09818`. 22

[214] Mahesh M. Thakkar, Rishi Sharma, and Pradeep Sahota. Alcohol disrupts sleep homeostasis. *Alcohol*, 49(4):299–310, 2015. 8, 113

[215] Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore, Gauri Ganjoo, Emmanuel Mignot, and James Zou. SleepFM: Multi-modal Representation Learning for Sleep Across Brain Activity, ECG and Respiratory Signals, 2024, `http://arxiv.org/abs/2405.17766`. 4, 104, 145

[216] Ståle Toften, Ståle Pallesen, Maria Hrozanova, Frode Moen, and Janne Grønli. Validation of sleep stage classification using non-contact radar technology and machine learning (Somnofy®). *Sleep Medicine*, 75:54–61, 2020. Publisher: Elsevier. 32

[217] Alexander Trumpp, Johannes Lohr, Daniel Wedekind, Martin Schmidt, Matthias Burghardt, Axel R. Heller, Hagen Malberg, and Sebastian Zaunseder. Camera-based photoplethysmography in an intraoperative setting. *BioMedical Engineering OnLine*, 17(1):33, 2018. 24, 26, 44, 61

[218] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization, 2017, `http://arxiv.org/abs/1607.08022`. 106

[219] U.S. Food and Drug Administration (FDA). De Novo Classification Request for Oxehealth Vital Signs, 2020, `https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN200019.pdf`. 5, 27, 31, 61, 63

[220] U.S. Food and Drug Administration (FDA). De Novo Classification Request for Gili Pro Biosensor, 2020, `https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN200038.pdf`. 61

[221] Caspar van der Woerd, Hans van Gorp, Sylvie Dujardin, Manuel Sastry, Humberto Garcia Caballero, Fokke van Meulen, Stef van den Elzen, Sebastiaan Overeem, and Pedro Fonseca. Studying sleep: towards the identification of hypnogram features that drive expert interpretation. *Sleep*, page zsad306, 2023. 142

[222] Mark van Gastel, Sander Stuijk, and Gerard de Haan. Motion Robust Remote-PPG in Infrared. *IEEE Transactions on Biomedical Engineering*, 62(5):1425–1433, 2015. 27

[223] Mark van Gastel, Sander Stuijk, Sebastiaan Overeem, Johannes P. van Dijk, Merel M. van Gilst, and Gerard de Haan. Camera-Based Vital Signs Monitoring During Sleep – A Proof of Concept Study. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1409–1418, 2021. 26, 38, 44, 61

[224] Hans van Gorp, Iris A M Huijben, Pedro Fonseca, Ruud J G van Sloun, Sebastiaan Overeem, and Merel M van Gilst. Certainty about uncertainty in sleep staging: a theoretical framework. *Sleep*, 45(8), 2022. 55, 60, 112, 114, 136

[225] Hans van Gorp, Merel M. van Gilst, Pedro Fonseca, Sebastiaan Overeem, and Ruud J.G. van Sloun. Modeling the Impact of Inter-Rater Disagreement on Sleep Statistics using Deep Generative Learning. *IEEE Journal of Biomedical and Health Informatics*, pages 1–12, 2023. 118

[226] Fokke B. van Meulen, Angela Grassi, Leonie van den Heuvel, Sebastiaan Overeem, Merel M. van Gilst, Johannes P. van Dijk, Henning Maass, Mark J. H. van Gastel, and Pedro Fonseca. Contactless Camera-Based Sleep Staging: The HealthBed Study. *Bioengineering*, 10(1):109, 2023. 4, 32, 39, 40, 49, 59, 62, 70

[227] Emilio Vanoli, Philip B. Adamson, null Ba-Lin, Gian D. Pinna, Ralph Lazzara, and William C. Orr. Heart Rate Variability During Specific Sleep Stages. *Circulation*, 91(7):1918–1922, 1995. Publisher: American Heart Association. 113

[228] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv:1706.03762 [cs]*, 2017. arXiv: 1706.03762. 3, 18, 19, 20, 21, 22, 30, 52, 57, 66, 81, 92, 119

[229] Christian Veauthier, Juliane Ryczewski, Sebastian Mansow-Model, Karen Otte, Bastian Kayser, Martin Glos, Christoph Schöbel, Friedemann Paul, Alexander U. Brandt, and Thomas Penzel. Contactless recording of sleep apnea and periodic leg movements by nocturnal 3-D-video and subsequent visual perceptive computing. *Scientific Reports*, 9(1):16812, 2019. 40

[230] Wim Verkruysse, Lars O. Svaasand, and J. Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics Express*, 16(26):21434–21445, 2008. 44

[231] Mauricio Villarroel, Alessandro Guazzi, João Jorge, Sara Davis, Peter Watkinson, Gabrielle Green, Asha Shenvi, Kenny McCormick, and Lionel Tarassenko. Continuous non-contact vital sign monitoring in neonatal intensive care unit. *Healthcare Technology Letters*, 1(3):87–91, 2014. 24, 26

[232] Mauricio Villarroel, Sitthichok Chaichulee, João Jorge, Sara Davis, Gabrielle Green, Carlos Arteta, Andrew Zisserman, Kenny McCormick, Peter Watkinson, and Lionel Tarassenko. Non-contact physiological monitoring of preterm infants in the Neonatal Intensive Care Unit. *npj Digital Medicine*, 2(1):1–18, 2019. 24, 61

[233] Mauricio Villarroel, João Jorge, David Meredith, Sheera Sutherland, Chris Pugh, and Lionel Tarassenko. Non-contact vital-sign monitoring of patients undergoing haemodialysis treatment. *Scientific Reports*, 10(1):18529, 2020. 24, 26, 31, 44

[234] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, pages I–511–I–518, Kauai, HI, USA, 2001. IEEE Comput. Soc. 25

[235] Tom Vogels, Mark Van Gastel, Wenjin Wang, and Gerard De Haan. Fully-Automatic Camera-Based Pulse-Oximetry During Sleep. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1430–14308, Salt Lake City, UT, USA, 2018. IEEE. 28

[236] Olivia Walch, Yitong Huang, Daniel Forger, and Cathy Goldstein. Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. *Sleep*, 42(12), 2019. 38, 54, 59

[237] Jiquan Wang, Sha Zhao, Haiteng Jiang, Yangxuan Zhou, Zhenghe Yu, Tao Li, Shijian Li, and Gang Pan. CareSleepNet: A Hybrid Deep Learning Network for Automatic Sleep Staging. *IEEE Journal of Biomedical and Health Informatics*, pages 1–14, 2024. 90

[238] Qiongyan Wang, Hanrong Cheng, and Wenjin Wang. Feasibility of Exploiting Physiological and Motion Features for Camera-based Sleep Staging: A Clinical Study. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–5, 2023. ISSN: 2694-0604. 26, 27, 61

[239] Qiongyan Wang, Hanrong Cheng, and Wenjin Wang. Video-PSG: An Intelligent Contactless Monitoring System for Sleep Staging. *IEEE Transactions on Biomedical Engineering*, 2024. Publisher: IEEE. 4

[240] Wenjin Wang and Albertus C. den Brinker. Algorithmic insights of camera-based respiratory motion extraction. *Physiological measurement*, 43(7):075004, 2022. 44, 59, 61

[241] Wenjin Wang, Albertus C. den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic Principles of Remote PPG. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2017. 44, 59, 61

[242] Wenjin Wang, Albertus C. den Brinker, and Gerard de Haan. Single-Element Remote-PPG. *IEEE Transactions on Biomedical Engineering*, 66(7):2032–2043, 2019. 25

[243] Wenjin Wang, Albertus C. den Brinker, and Gerard de Haan. Discriminative Signatures for Remote-PPG. *IEEE Transactions on Biomedical Engineering*, 67(5): 1462–1473, 2020. 27

[244] Wenjin Wang, Luc Vosters, and Albertus C. den Brinker. Continuous-Spectrum Infrared Illuminator for Camera-PPG in Darkness. *Sensors*, 20(11):3044, 2020. 26, 27

[245] William H. Ward, Fernando Lambreton, Neha Goel, Jian Q. Yu, and Jeffrey M. Farma. Clinical Presentation and Staging of Melanoma. In *Cutaneous Melanoma: Etiology and Therapy*. Codon Publications, Brisbane (AU), 2017. 72

[246] Johanna Wilde-Frenz and Hartmut Schulz. Rate and Distribution of Body Movements during Sleep in Humans. *Perceptual and Motor Skills*, 56(1):275–283, 1983. 31, 59, 66

[247] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM transactions on graphics (TOG)*, 31(4):1–8, 2012. Publisher: ACM New York, NY, USA. 25

[248] Bernice M Wulterkens, Pedro Fonseca, Lieke W A Hermans, Marco Ross, Andreas Cerny, Peter Anderer, Xi Long, Johannes P van Dijk, Nele Vandenbussche, Sigrid Pillen, Merel M van Gilst, and Sebastiaan Overeem. It is All in the Wrist: Wearable Sleep Staging in a Clinical Population versus Reference Polysomnography. *Nature and Science of Sleep*, 13:885–897, 2021. 38

[249] Marina Wunderlin, Marc Alain Züst, Kristoffer Daniel Fehér, Stefan Klöppel, and Christoph Nissen. The role of slow wave sleep in the development of dementia and its potential for preventative interventions. *Psychiatry Research: Neuroimaging*, 306:111178, 2020. 2

[250] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pages 10524–10533. PMLR, 2020. 81

[251] Shuting Xu, Oliver Faust, Silvia Seoni, Subrata Chakraborty, Prabal Datta Barua, Hui Wen Loh, Heather Elphick, Filippo Molinari, and U. Rajendra Acharya. A review of automated sleep disorder detection. *Computers in Biology and Medicine*, 150:106100, 2022. 144

[252] Biao Xue, Boya Deng, Hong Hong, Zhiyong Wang, Xiaohua Zhu, and David Dagan Feng. Non-contact sleep stage detection using canonical correlation analysis of respiratory sound. *IEEE Journal of Biomedical and Health Informatics*, 24(2): 614–625, 2019. Publisher: IEEE. 32

[253] Yuzhe Yang, Xin Liu, Jiang Wu, Silviu Borac, Dina Katabi, Ming-Zher Poh, and Daniel McDuff. SimPer: Simple Self-Supervised Learning of Periodic Targets, 2023. 26, 61

[254] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks?, 2014, `http://arxiv.org/abs/1411.1792`. 23

[255] Magdy Younes, Jill Raneri, and Patrick Hanly. Staging Sleep in Polysomnograms: Analysis of Inter-Scorer Variability. *Journal of Clinical Sleep Medicine*, 12(06): 885–894. Publisher: American Academy of Sleep Medicine. 13

[256] Terry Young, 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. *WMJ: official publication of the State Medical Society of Wisconsin*, 108(5):246–249, 2009. 93, 154

[257] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip Torr, and Guoying Zhao. PhysFormer: Facial Video-based Physiological Measurement with Temporal Difference Transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4176–4186, New Orleans, LA, USA, 2022. IEEE. 26, 61

[258] Hang Yuan, Shing Chan, Andrew P. Creagh, Catherine Tong, Aidan Acquah, David A. Clifton, and Aiden Doherty. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *npj Digital Medicine*, 7(1): 1–10, 2024. Publisher: Nature Publishing Group. 4

[259] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing. 18

[260] Guo-Qiang Zhang, 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. *Journal of the American Medical Informatics Association: JAMIA*, 25(10):1351–1358, 2018. 13, 53, 62, 86, 93, 153, 155

[261] Hangyu Zhu, Wei Zhou, Cong Fu, Yonglin Wu, Ning Shen, Feng Shu, Huan Yu, Wei Chen, and Chen Chen. MaskSleepNet: A Cross-Modality Adaptation Neural Network for Heterogeneous Signals Processing in Sleep Staging. *IEEE Journal of Biomedical and Health Informatics*, 27(5):2353–2364, 2023. 90

[262] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 23

[263] Dan Ziegler, G. Laux, K. Dannehl, M. Spüler, H. Mühlen, P. Mayer, and F.a. Gries. Assessment of Cardiovascular Autonomic Function: Age-related Normal Ranges and Reproducibility of Spectral Analysis, Vector Analysis, and Standard Tests of Heart Rate Variation and Blood Pressure Responses. *Diabetic Medicine*, 9(2):166–175, 1992. 50, 72, 77